A PROXIMAL METHOD FOR COMPOSITE MINIMIZATION

A. S. LEWIS* AND S. J. WRIGHT†

Abstract. We consider minimization of functions that are compositions of prox-regular functions with smooth vector functions. A wide variety of important optimization problems can be formulated in this way. We describe a subproblem constructed from a linearized approximation to the objective and a regularization term, investigating the properties of local solutions of this subproblem and showing that they eventually identify a manifold containing the solution of the original problem. We propose an algorithmic framework based on this subproblem and prove a global convergence result.

Key words. prox-regular functions, polyhedral convex functions, sparse optimization, global convergence, active constraint identification

AMS subject classifications. 49M37, 90C30

1. Problem Statement.

We consider minimization problems of the form

$$\min_x \; h\big(c(x)\big), \tag{1.1}$$

where the inner function $c : \mathbb{R}^n \to \mathbb{R}^m$ is smooth. On the other hand, the outer function $h : \mathbb{R}^m \to [-\infty, +\infty]$ may be nonsmooth, but is usually convex, and in some way structured: it is often even polyhedral. Assuming that h is sufficiently well-structured to allow us to solve, relatively easily, subproblems of the form

$$\min_d \; h\big(\Phi(d)\big) + \frac{\mu}{2}|d|^2, \tag{1.2}$$

for *affine* maps Φ and scalars $\mu > 0$ (where $|\cdot|$ denotes the Euclidean norm throughout the paper), we design and analyze a "proximal" method for the problem (1.1). More precisely, we consider an algorithmic framework in which a *proximal linearized subproblem* of the form (1.2) is solved at each iteration to define a first approximation to a step. If the function h is sufficiently well-structured—an assumption we make concrete using "partial smoothness," a generalization of the idea of an active set in nonlinear programming—we may then be able to enhance the step, possibly with the use of higher-order derivative information.

Many important problems in the form (1.1) involve finite convex functions h. Our development explores, nonetheless, to what extent the underlying theory for the proposed algorithm extends to more general functions. Specifically, we broaden the class of allowable functions h in two directions:

- h may be extended-valued, allowing constraints that must be enforced;
- we weaken the requirement of convexity to "prox-regularity".

This broader framework involves extra technical overhead, but we point out throughout how the development simplifies in the case of continuous convex h, and in particular polyhedral h.

Let us fix some notation. We consider a local solution (or, more generally, critical point) $\bar{x}$ for the problem (1.1), and let $\bar{c} := c(\bar{x})$. (Our assumption that the function c

*ORIE, Cornell University, Ithaca, NY 14853, U.S.A. aslewis@orie.cornell.edu people.orie.cornell.edu/~aslewis. Research supported in part by National Science Foundation Grant DMS-0806057.

†Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706. swright@cs.wisc.edu pages.cs.wisc.edu/~swright. Research supported in part by National Science Foundation Grant 0430504.

is everywhere defined is primarily for notational simplicity: restricting our analysis to a neighborhood of $\bar{x}$ is straightforward.) The criticality condition is $0 \in \partial(h \circ c)(\bar{x})$, where ∂ denotes the subdifferential. As we discuss below, under reasonable conditions, a chain rule then implies existence of a vector $\bar{v}$ such that

$$\bar{v} \in \partial h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*), \tag{1.3}$$

where $\nabla c(\bar{x}) : \mathbb{R}^n \to \mathbb{R}^m$ is the derivative of c at $\bar{x}$ and * denotes the adjoint map. In typical examples, we can interpret the vector(s) $\bar{v}$ as Lagrange multipliers, as we discuss below.

We prove results of three types:

1. When the current point x is near the critical point $\bar{x}$, the proximal linearized subproblem (1.2) has a local solution d of size $O(|x - \bar{x}|)$. By projecting the point $x + d$ onto the inverse image under the map c of the domain of the function h, we can obtain a step that reduces the objective (1.1).
2. Under reasonable conditions, when x is close to $\bar{x}$, if h is "partly smooth" at $\bar{c}$ relative to a certain manifold $\mathcal{M}$ (a generalization of the surface defined by the active constraints in classical nonlinear programming), then the algorithm "identifies" $\mathcal{M}$: The solution d of the subproblem (1.2) has $\Phi(d) \in \mathcal{M}$.
3. A global convergence result for an algorithm based on (1.2).

1.1. Definitions. We begin with some important definitions. We write $\bar{\mathbb{R}}$ for the extended reals $[-\infty, +\infty]$, and consider a function $h : \mathbb{R}^m \to \bar{\mathbb{R}}$. The notion of the subdifferential of h at a point $\bar{c} \in \mathbb{R}^m$, denoted $\partial h(\bar{c})$, provides a powerful unification of the classical gradient of a smooth function, and the subdifferential from convex analysis. It is a set of generalized gradient vectors, coinciding exactly with the classical convex subdifferential [33] when h is lower semicontinuous and convex, and equalling $\{\nabla h(\bar{c})\}$ when h is $\mathcal{C}^1$ around $\bar{c}$. For the formal definition, and others from variational analysis, the texts [34] and [27] are good sources.

An elegant framework for unifying smooth and convex analysis is furnished by the notion of "prox-regularity" [29]. Geometrically, the idea is rather natural: a set in $S \subset \mathbb{R}^m$ is *prox-regular* at a point $s \in S$ if every point near s has a unique nearest point in S (using the Euclidean distance). In particular, closed convex sets are prox-regular at every point. A finite collection of $\mathcal{C}^2$ equality and inequality constraints defines a set that is prox-regular at any point where the gradients of the active constraints are linearly independent.

A function $h : \mathbb{R}^m \to \bar{\mathbb{R}}$ is *prox-regular* at a point $\bar{c}$ if $h(\bar{c})$ is finite and the epigraph

$$\mathrm{epi}\, h \;:=\; \big\{(c, r) \in \mathbb{R}^m \times \mathbb{R} : r \geq h(c)\big\}$$

is prox-regular at the point $\big(\bar{c}, h(\bar{c})\big)$. In particular, both convex and $\mathcal{C}^2$ functions are prox-regular wherever they are defined.

A general class of prox-regular functions common in engineering applications are "lower $\mathcal{C}^2$" (see Rockafellar and Wets [34]). A function $h : \mathbb{R}^m \to \mathbb{R}$ is *lower* $\mathcal{C}^2$ around a point $\bar{c} \in \mathbb{R}^m$ if h has the local representation

$$h(c) = \max_{t \in T} f(c, t) \;\text{ for } c \in \mathbb{R}^m \text{ near } \bar{c},$$

for some function $f : \mathbb{R}^m \times T \to \mathbb{R}$, where the space T is compact and the quantities $f(c, t)$, $\nabla_c f(c, t)$, and $\nabla^2_{cc} f(c, t)$ all depend continuously on (c, t). A simple equivalent property, useful in theory though harder to check in practice, is that h has the form

$g - \kappa|\cdot|^2$ around the point $\bar{c}$ for some continuous convex function g and some constant κ.

The original definition of prox-regularity given by Poliquin and Rockafellar [29] involved the subdifferential, as follows. For the equivalence with the geometric definition above, see Poliquin, Rockafellar, and Thibault [30].

DEFINITION 1.1. *A function $h : \mathbb{R}^m \to \bar{\mathbb{R}}$ is* prox-regular at a point $\bar{c} \in \mathbb{R}^m$ for a subgradient $\bar{v} \in \partial h(\bar{c})$ *if h is finite at $\bar{c}$, locally lower semicontinuous around $\bar{c}$, and there exists $\rho > 0$ such that*

$$h(c') \geq h(c) + \langle v, c' - c\rangle - \frac{\rho}{2}|c' - c|^2$$

whenever points $c, c' \in \mathbb{R}^m$ are near $\bar{c}$ with the value $h(c)$ near the value $h(\bar{c})$ and for every subgradient $v \in \partial h(c)$ near $\bar{v}$. Further, h is prox-regular at $\bar{c}$ *if it is prox-regular at $\bar{c}$ for every $\bar{v} \in \partial h(\bar{c})$.*

Note in particular that if h is prox-regular at $\bar{c}$, we have that, for every $\bar{v} \in \partial h(\bar{c})$, there exists $\rho > 0$ such that

$$h(c') \geq h(\bar{c}) + \langle \bar{v}, c' - \bar{c}\rangle - \frac{\rho}{2}|c' - \bar{c}|^2, \tag{1.4}$$

whenever c' is near $\bar{c}$. (Set $c = \bar{c}$ in the definition above.)

A weaker property than the prox-regularity of a function h is "subdifferential regularity," a concept easiest to define in the case in which h is Lipschitz. In this case, h is almost everywhere differentiable: it is *subdifferentially regular* at a point $\bar{c} \in \mathbb{R}^m$ if its classical directional derivative for every direction $d \in \mathbb{R}^m$ equals

$$\limsup_{c \to \bar{c}} \langle \nabla h(c), d\rangle,$$

where the lim sup is taken over points c where h is differentiable. Clearly, $\mathcal{C}^1$ functions have this property; continuous convex functions also have it. For nonlipschitz functions the notion is less immediate to define (see Rockafellar and Wets [34]), but it holds for lower semicontinuous, convex functions (see [34, Example 7.27]) and more generally for prox-regular functions.

We next turn to the idea of "partial smoothness" introduced by Lewis [22], a variational-analytic formalization of the notion of the active set in classical nonlinear programming. The notion we describe here is, more precisely, "$\mathcal{C}^2$-partial smoothness": see Hare and Lewis [16, Definition 2.3]. In the definition below, a set $\mathcal{M} \subset \mathbb{R}^m$ is a *manifold about* a point $\bar{c} \in \mathcal{M}$ if it can be described locally by a collection of smooth equations with linearly independent gradients: more precisely, there exists a map $F : \mathbb{R}^m \to \mathbb{R}^k$ that is $\mathcal{C}^2$ around $\bar{c}$ with $\nabla F(\bar{c})$ surjective and such that points $c \in \mathbb{R}^m$ near $\bar{c}$ lie in $\mathcal{M}$ if and only if $F(c) = 0$. The classical *normal space* to $\mathcal{M}$ at $\bar{c}$, denoted $N_{\mathcal{M}}(\bar{c})$ is then just the range of $\nabla F(\bar{c})^*$.

DEFINITION 1.2. *A function $h : \mathbb{R}^m \to \bar{\mathbb{R}}$ is* partly smooth *at a point $\bar{c} \in \mathbb{R}^m$ relative to a set $\mathcal{M} \subset \mathbb{R}^m$ containing $\bar{c}$ if $\mathcal{M}$ is a manifold about $\bar{c}$ and the following properties hold:*

(i) (Smoothness) *The restricted function $h\,|_{\mathcal{M}}$ is $\mathcal{C}^2$ near $\bar{c}$;*
(ii) (Regularity) *h is subdifferentially regular at all points $c \in \mathcal{M}$ near $\bar{c}$, with $\partial h(c) \neq \emptyset$;*
(iii) (Sharpness) *The affine span of $\partial h(\bar{c})$ is a translate of $N_{\mathcal{M}}(\bar{c})$;*
(iv) (Sub-continuity) *The set-valued mapping $\partial h : \mathcal{M} \rightrightarrows \mathbb{R}^m$ is continuous at $\bar{c}$.*

We refer to $\mathcal{M}$ as the active manifold.

A set $S \subset \mathbb{R}^m$ is *partly smooth* at a point $\bar{c} \in S$ relative to a manifold $\mathcal{M}$ if its indicator function,

$$\delta_S(c) = \begin{cases} 0 & (c \in S) \\ +\infty & (c \notin S), \end{cases}$$

is partly smooth at $\bar{c}$ relative to $\mathcal{M}$. Again we refer to $\mathcal{M}$ as the *active manifold*.

We denote by $P_S(v)$ the usual Euclidean projection of a vector $v \in \mathbb{R}^m$ onto a closed set $S \subset \mathbb{R}^m$. The *distance* between x and the set S is

$$\operatorname{dist}(x, S) = \inf_{y \in S} |x - y|.$$

We use $B_\epsilon(x)$ to denote the closed Euclidean ball of radius ϵ around a point x.

2. Examples. The framework (1.1) admits a wide variety of interesting problems, as we show in this section.

2.1. Approximation Problems.

EXAMPLE 2.1 (least squares, ℓ_1, and Huber approximation). *The formulation (1.1) encompasses both the usual (nonlinear) least squares problem if we define $h(\cdot) = |\cdot|^2$, and the ℓ_1 approximation problem if we define $h(\cdot) = |\cdot|_1$, the ℓ_1-norm. Another popular robust loss function is the Huber function defined by $h(c) = \sum_{i=1}^m \phi(c_i)$, where*

$$\phi(c_i) = \begin{cases} \frac{1}{2}c_i^2 & (|c_i| \le T) \\ Tc_i - \frac{1}{2}T^2 & (|c_i| > T). \end{cases}$$

EXAMPLE 2.2 (sum of Euclidean norms). *Given a collection of smooth vector functions $g_i : \mathbb{R}^n \to \mathbb{R}^{m_i}$, for $i = 1, 2, \ldots, t$, consider the problem*

$$\min_x \sum_{i=1}^t |g_i(x)|.$$

We can place such problems in the form (1.1) by defining the smooth vector function $c : \mathbb{R}^n \to \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \cdots \times \mathbb{R}^{m_t}$ by $c = (g_1, g_2, \ldots, g_t)$, and the nonsmooth function $h : \mathbb{R}^{m_2} \times \cdots \times \mathbb{R}^{m_t} \to \mathbb{R}$ by

$$h(g_1, g_2, \ldots, g_t) = \sum_{i=1}^t |g_i|.$$

2.2. Problems from Nonlinear Programming. Next, we consider examples motivated by penalty functions for nonlinear programming.

EXAMPLE 2.3 (ℓ_1 penalty function). *Consider the following nonlinear program:*

$$\begin{aligned} \min\ & f(x) \\ \text{subject to}\ & g_i(x) = 0 \quad (1 \le i \le j), \\ & g_i(x) \le 0 \quad (j \le i \le k), \\ & x \in X, \end{aligned} \tag{2.1}$$

where the polyhedron $X \subset \mathbb{R}^n$ describes constraints on the variable x that are easy to handle directly. The ℓ_1 penalty function formulation is

$$\min_{x \in X} f(x) + \nu \sum_{i=1}^{j} |g_i(x)| + \nu \sum_{i=j+1}^{k} \max\big(0, g_i(x)\big), \tag{2.2}$$

where $\nu > 0$ is a scalar parameter. We can express this problem in the form (1.1) by defining the smooth vector function

$$c(x) \;=\; \Big(f(x)\,,\, \big(g_i(x)\big)_{i=1}^{k}\,,\, x\Big) \;\in\; \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^n$$

and the extended polyhedral convex function $h : \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^n \to \bar{\mathbb{R}}$ by

$$h(f,g,x) = \begin{cases} f + \nu \sum_{i=1}^{j} |g_i| + \nu \sum_{i=j+1}^{k} \max(0, g_i) & (x \in X) \\ +\infty & (x \notin X). \end{cases}$$

A generalization of Example 2.3 in which h is a finite polyhedral function was the focus of much research in the 1980s. We consider this case further in Section 3 and use it again during the paper to illustrate the theory that we develop.

2.3. Regularized Minimization Problems. A large family of instances of (1.1) arises in the area of regularized minimization, where the minimization problem has the following general form:

$$\min_{x} f(x) + \tau |x|_* \tag{2.3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth objective, while $|x|_*$ is a continuous, nonnegative, usually nonsmooth function, and τ is a nonnegative *regularization parameter*. Such formulations arise when we seek an approximate minimizer of f that is "simple" in some sense; the purpose of the second term $|x|_*$ is to promote this simplicity property. Larger values of τ tend to produce solutions x that are simpler, but less accurate as minimizers of f. The problem (2.3) can be put into the framework (1.1) by defining

$$c(x) = \begin{bmatrix} f(x) \\ x \end{bmatrix} \in \mathbb{R}^{n+1}, \qquad h(f,x) = f + \tau |x|_*. \tag{2.4}$$

We list now some interesting cases of (2.3).

EXAMPLE 2.4 (ℓ_1-regularized minimization). *The choice $|\cdot|_* = |\cdot|_1$ in (2.3) tends to produce solutions x that are* sparse*, in the sense of having relatively few nonzero components. Larger values of τ tend to produce sparser solutions. Compressed sensing is a particular area of current interest, in which the objective f is typically a least-squares function $f(x) = (1/2)|Ax - b|^2$; see [6] for a recent survey. Regularized least-squares problems (or equivalent constrained-optimization formulations) are also encountered in statistics; see for example the LASSO [38] and LARS [12] procedures, and basis pursuit [7].*

A related application is regularized logistic regression, where again $|\cdot|_* = |\cdot|_1$, but f is (the negative of) an a posteriori log likelihood function. In the setup of [36], x contains the coefficients of a basis expansion of a log-odds ratio function, where each basis function is a function of the feature vector. The objective f is the (negative) log

likelihood function obtained by matching this data to a set of binary labels. In this case, f is convex but highly nonlinear. The regularization term causes the solution to have few nonzero coefficients, so the formulation identifies the most important basis functions for predicting the observed labels.

Another interesting class of regularized minimization problems arises in *matrix completion*, where we seek an $m \times n$ matrix X of smallest rank that is consistent with given knowledge of various linear combinations of the elements of X; see [5, 31, 4]. Much as the ℓ_1 of a vector x is used as a surrogate for cardinality of x in the formulations of Example 2.4, the *nuclear norm* is used as a surrogate for the rank of X in formulations of the matrix completion problem. The nuclear norm $|X|_*$ is defined as the sum of singular values of X, and we have the following specialization of (2.3):

$$\min_{X \in \mathbb{R}^{m \times n}} \tfrac{1}{2}|\mathcal{A}(X) - b|^2 + \tau |X|_*, \tag{2.5}$$

where $\mathcal{A}$ denotes a linear operator from $\mathbb{R}^{m \times n}$ to $\mathbb{R}^p$, and $b \in \mathbb{R}^p$ is the observation vector. Note that the nuclear norm is a continuous and convex function of X.

2.4. Nonconvex Problems. Each of the examples above involves a convex outer function h. In principle, however, the techniques we develop here also apply to a variety of nonconvex functions. The next example includes some simple illustrations.

EXAMPLE 2.5 (problems involving quadratics). *Given a general quadratic function $f : \mathbb{R}^p \to \mathbb{R}$ (possibly nonconvex) and a smooth function $c_1 : \mathbb{R}^n \to \mathbb{R}^p$, consider the problem $\min_x f\big(c_1(x)\big)$. This problem trivially fits into the framework (1.1), and the function f, being $\mathcal{C}^2$, is everywhere prox-regular. The subproblems (1.2), for sufficiently large values of the parameter μ, simply amount to solving a linear system.*

More generally, given another general quadratic function $g : \mathbb{R}^q \to \mathbb{R}$, and another smooth function $c_2 : \mathbb{R}^n \to \mathbb{R}^q$, consider the problem

$$\min_{x \in \mathbb{R}^n} f\big(c_1(x)\big) \quad \text{subject to} \quad g\big(c_2(x)\big) \le 0.$$

We can express this problem in the form (1.1) by defining the smooth vector function $c = (c_1, c_2)$ and defining an extended-valued nonconvex function

$$h(c_1, c_2) = \begin{cases} f(c_1) & (g(c_2) \le 0) \\ +\infty & (g(c_2) > 0). \end{cases}$$

The epigraph of h is

$$\big\{(c_1, c_2, t) : g(c_2) \le 0,\ t \ge f(c_1)\big\},$$

a set defined by two smooth inequality constraints: hence h is prox-regular at any point (c_1, c_2) satisfying $g(c_2) \le 0$ and $\nabla g(c_2) \ne 0$. The resulting subproblems (1.2) are all in the form of the standard trust-region subproblem, and hence relatively straightforward to solve quickly.

As one more example, consider the case when the outer function h is defined as the maximum of a finite collection of quadratic functions (possibly nonconvex): $h(x) = \max\{f_i(x) : i = 1, 2, \ldots, k\}$. We can write the subproblems (1.2) in the form

$$\min \Big\{ t : t \ge f_i\big(\Phi(d)\big) + \frac{\mu}{2}|d|^2,\ d \in \mathbb{R}^m,\ t \in \mathbb{R},\ i = 1, 2, \ldots, k \Big\}.$$

where the map Φ is affine. For sufficiently large values of the parameter μ, this quadratically-constrained convex quadratic program can in principle be solved efficiently by an interior point method.

To conclude, we consider two more applied nonconvex examples. The first is due to Mangasarian [23] and is used by Jokar and Pfetsch [18] to find sparse solutions of underdetermined linear equations. The formulation of [18] can be stated in the form (2.3) where the regularization function $|\cdot|_*$ has the form

$$|x|_* = \sum_{i=1}^{n}(1 - e^{-\alpha|x_i|})$$

for some parameter $\alpha > 0$. It is easy to see that this function is nonconvex but prox-regular, and nonsmooth only at $x_i = 0$.

Zhang et al. [43] use a similar regularization function of the form (2.3) that behaves like the ℓ_1 norm near the origin and transitions (via a concave quadratic) to a constant for large loss values. Specifically, we have $|\cdot|_* = \sum_{i=1}^{n}\phi(x_i)$, where

$$\phi(x_i) = \begin{cases} \lambda|x_i| & (|x_i| \le \lambda) \\ -(|x_i|^2 - 2a\lambda|x_i| + \lambda^2)/\big(2(a-1)\big) & (\lambda < |x_i| \le a\lambda) \\ (a+1)\lambda^2/2 & (|x_i| > a\lambda). \end{cases}$$

Here $\lambda > 0$ and $a > 1$ are tuning parameters.

3. The Finite Polyhedral Case. As we have remarked, a classical example of our model problem $\min_x h\big(c(x)\big)$ is the case in which the outer function h is finite and polyhedral:

$$h(c) = \max_{i\in I}\{\langle h_i, c\rangle + \beta_i\} \tag{3.1}$$

for some given vectors $h_i \in \mathbb{R}^m$ and scalars β_i, where the index i runs over some finite set I. We use this case to illustrate much of the basic theory we develop.

Assume the map $c\colon \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^1$ around a critical point $\bar{x} \in \mathbb{R}^n$ for the composite function $h \circ c$, and let $\bar{c} = c(\bar{x})$. Define the set of "active" indices

$$\bar{I} \;=\; \operatorname{argmax}\big\{\langle h_i, \bar{c}\rangle + \beta_i : i \in I\big\}.$$

Then, denoting convex hulls by conv, we have

$$\partial h(\bar{c}) = \operatorname{conv}\{h_i : i \in \bar{I}\}.$$

Hence the basic criticality condition (1.3) becomes the existence of a vector $\lambda \in \mathbb{R}^{\bar{I}}$ satisfying

$$\lambda \ge 0 \quad \text{and} \quad \sum_{i\in\bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{3.2}$$

The vector $\bar{v}$ is then $\sum_{i\in\bar{I}} \lambda_i h_i$.

Compare this with the classical nonlinear programming framework, which is

$$\begin{aligned} &\min\ t \\ &\text{subject to} \quad \langle h_i, c(x)\rangle + \beta_i + t \le 0 \quad (i \in I) \\ &\qquad\qquad\qquad\qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}. \end{aligned} \tag{3.3}$$

At the point $\big(\bar{x}, -h(\bar{c})\big)$, the conditions (3.2) are just the standard first-order optimality conditions, with Lagrange multipliers λ_i. The fact that the vector $\bar{v}$ in the criticality condition (1.3) is closely identified with λ via the relationship $\bar{v} = \sum_{i \in \bar{I}} \lambda_i h_i$ motivates our terminology "multiplier vector".

We return to this example repeatedly in what follows.

4. The Proximal Linearized Subproblem. In this work we consider an algorithmic framework based on solution of a proximal linearized subproblem of the form (1.2) at each iteration. We focus on the case in which $\Phi(d)$ is the Taylor-series linearization of c around the current iterate x, which yields the following subproblem:

$$\min_d \; h_{x,\mu}(d) := h\big(c(x) + \nabla c(x)d\big) + \frac{\mu}{2}|d|^2, \tag{4.1}$$

where $\mu > 0$ is a parameter and the linear map $\nabla c(x) : \mathbb{R}^n \to \mathbb{R}^m$ is the derivative of the map c at x (representable by the $m \times n$ Jacobian matrix).

For simplicity, consider first a function $h : \mathbb{R}^m \to (-\infty, +\infty]$ that is convex and lower semicontinuous. Assuming that the vector $c(x) + \nabla c(x)d$ lies in the domain of h for some step $d \in \mathbb{R}^n$, the subproblem (4.1) involves minimizing a strictly convex function with nonempty compact level sets, and thus has a unique solution $d = d(x)$. If we assume slightly more — that $c(x) + \nabla c(x)d$ lies in relative interior of the domain of h for some d (as holds obviously if h is continuous at $c(x)$), a standard chain rule from convex analysis implies that $d = d(x)$ is the unique solution of the following inclusion:

$$\nabla c(x)^* v + \mu d = 0, \;\; \text{for some } v \in \partial h\big(c(x) + \nabla c(x)d\big). \tag{4.2}$$

When h is just prox-regular rather than convex, under reasonable conditions (see below), the subproblem (4.1) still has a unique local solution close to zero, for μ sufficiently large, which is characterized by property (4.2).

For regularized minimization problems of the form (2.3), the subproblem (4.1) has the form

$$\min_d \; f(x) + \langle \nabla f(x), d\rangle + \frac{\mu}{2}|d|^2 + \tau|x + d|_*. \tag{4.3}$$

An equivalent formulation can be obtained by shifting the objective and making the change of variable $z := x + d$:

$$\min_z \; \frac{\mu}{2}|z - y|^2 + \tau|z|_*, \qquad \text{where} \quad y = x - \frac{1}{\mu}\nabla f(x). \tag{4.4}$$

When the regularization function $|\cdot|_*$ is separable in the components of x, as when $|\cdot|_* = |\cdot|_1$ or $|\cdot| = |\cdot|_2^2$, this problem can be solved in $O(n)$ operations. (Indeed, this fact is key to the efficiency of methods based on these subproblems in compressed sensing applications.) For the case $|\cdot|_* = |\cdot|_1$, if we set $\alpha = \tau/\mu$, the solution of (4.4) is

$$z_i = \begin{cases} 0 & (|y_i| \le \alpha) \\ y_i - \alpha & (y_i > \alpha) \\ y_i + \alpha & (y_i < -\alpha). \end{cases} \tag{4.5}$$

The operation specified by (4.5) is known commonly as the "shrink operator."

For the matrix completion formulation (2.5), the formulation (4.4) of the subproblem becomes

$$\min_{Z\in\mathbb{R}^{m\times n}} \frac{\mu}{2}|Z-Y|_F^2+\tau|Z|_*, \tag{4.6}$$

where $|\cdot|_F$ denotes the Frobenius norm of a matrix and

$$Y = X - \frac{1}{\mu}\mathcal{A}^*[\mathcal{A}(X)-b]. \tag{4.7}$$

It is known (see for example [4]) that (2.5) can be solved by using the singular-value decomposition of Y. Writing $Y = U\Sigma V^T$, where U and V are orthogonal and $\Sigma = \mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_{\min(m,n)})$, we have $Z = U\Sigma_{\tau/\mu}V^T$, where the diagonals of $\Sigma_{\tau/\mu}$ are $\max(\sigma_i-\tau/\mu,0)$ for $i=1,2,\ldots,\min(m,n)$. In essence, we apply the shrink operator to the singular values of Y, and reconstruct Z by using the orthogonal matrices U and V from the decomposition of Y.

5. Related Work. We discuss here some connections of our approach with existing literature.

We begin by considering various approaches when the outer function h is finite and polyhedral. One closely related work is by Fletcher and Sainz de la Maza [13], who discuss an algorithm for minimization of the ℓ_1-penalty function (2.2) for the nonlinear optimization problem (2.1). The first step of their method at each iteration is to solve a linearized trust-region problem which can be expressed in our general notation as follows:

$$\min_d h\big(c(x)+\nabla c(x)d\big) \ \text{ subject to } \ |d|\le\rho, \tag{5.1}$$

where ρ is some trust-region radius. Note that this subproblem is closely related to our linearized subproblem (4.1) when the Euclidean norm is used to define the trust region. However, the ℓ_∞ norm is preferred in [13], as the subproblem (5.1) can then be expressed as a linear program. The algorithm in [13] uses the solution of (5.1) to estimate the active constraint manifold, then computes a step that minimizes a model of the Lagrangian function for (2.1) while fixing the identified constraints as equalities. A result of active constraint identification is proved ([13, Theorem 2.3]); this result is related to our Theorems 6.12 and 7.5 below.

Byrd et al. [3] describe a successive linear-quadratic programming method, based on [13], which starts with solution of the linear program (5.1) (with ℓ_∞ trust region) and uses it to define an approximate Cauchy point, then approximately solves an equality-constrained quadratic program (EQP) over a different trust region to enhance the step. This algorithm is implemented in the KNITRO package for nonlinear optimization as the KNITRO-ACTIVE option.

Friedlander et al. [14] solve a problem of the form (4.1) for the case of nonlinear programming, where h is the sum of the objective function f and the indicator function for the equalities and the inequalities defining the feasible region. The resulting step can be enhanced by solving an EQP.

Other related literature on composite nonsmooth optimization problems with general finite polyhedral convex functions (Section 3) includes the papers of Yuan [41, 42] and Wright [39]. The approaches in [42, 39] solve a linearized subproblem like (5.1), from which an analog of the "Cauchy point" for trust-region methods in smooth unconstrained optimization can be calculated. This calculation involves a line search

along a piecewise quadratic function and is therefore more complicated than the calculation in [13], but serves a similar purpose, namely as the basis of an acceptability test for a step obtained from a higher-order model.

For general outer functions h, the theory is more complex. An early approach to regularized minimization problems of the form (2.3) for a lower semicontinuous convex function $|\cdot|_*$ is due to Fukushima and Mine [15]: they calculate a trial step at each iteration by solving the linearized problem (4.3).

The case when the map c is simply the identity has a long history. The iteration $x_{k+1} = x_k + d_k$, where d_k minimizes the function $d \mapsto h(x_k + d) + \frac{\mu}{2}|d|^2$, is the well-known *proximal point method*. For lower semicontinuous convex functions h, convergence was proved by Martinet [24] and generalized by Rockafellar [32]. For nonconvex h, a good survey up to 1998 is by Kaplan and Tichatschke [19]. Pennanen [28] took an important step forward, showing in particular that if the graph of the subdifferential ∂h agrees locally with the graph of the inverse of a Lipschitz function (a condition verifiable using second-order properties including prox-regularity—see Levy [21, Cor. 3.2]), then the proximal point method converges linearly if started nearby and with regularization parameter μ bounded away from zero. This result was foreshadowed in much earlier work of Spingarn [37], who gave conditions guaranteeing local linear convergence of the proximal point method for a function h that is the sum of lower semicontinuous convex function and a $C2$ function, conditions which furthermore hold "generically" under perturbation by a linear function. Inexact variants of Pennanen's approach are discussed by Iusem, Pennanen, and Svaiter [17] and Combettes and Pennanen [8]. In this current work, we make no attempt to build on this more sophisticated theory, preferring a more direct and self-contained approach.

The issue of identification of the face of a constraint set on which the solution of a constrained optimization problem lies has been the focus of numerous other works. Some papers show that the projection of the point $x - \sigma\nabla f(x)$ onto the feasible set (for some fixed $\sigma > 0$) lies on the same face as the solution $\bar{x}$, under certain nondegeneracy assumptions on the problem and geometric assumptions on the feasible set. Identification of so-called quasi-polyhedral faces of convex sets was described by Burke and Moré [2]. An extension to the nonconvex case is provided by Burke [1], who considers algorithms that work with linearizations of the constraints describing the feasible set. Wright [40] considers surfaces of a convex set that can be parametrized by a smooth algebraic mapping, and shows how algorithms of gradient projection type can identify such surfaces once the iterates are sufficiently close to a solution. Lewis [22] and Hare and Lewis [16] extend these results to the nonconvex, nonsmooth case by using concepts from nonsmooth analysis, including partly smooth functions and prox-regularity. In their setting, the concept of a identifiable face of a feasible set becomes a certain type of manifold with respect to which h is partly smooth (see Definition 1.2 above). Their main results give conditions under which the active manifold is identified from within a neighborhood of the solution.

Another line of relevant work is associated with the $\mathcal{VU}$ theory introduced by Lemaréchal, Oustry, and Sagastizábal [20] and subsequently elaborated by these and other authors. The focus is on minimizing convex functions $f(x)$ that, again, are partly smooth — smooth ("U-shaped") along a certain manifold through the solution $\bar{x}$, but nonsmooth ("V-shaped") in the transverse directions. Mifflin and Sagastizábal [35] discuss the "fast track," which is essentially the manifold containing the solution $\bar{x}$ along which the objective is smooth. Similarly to [13], they are interested in algorithms that identify the fast track and then take a minimization step for a

certain Lagrangian function along this track. It is proved in [35, Theorem 5.2] that under certain assumptions, when x is near $\bar{x}$, the proximal point $x + d$ obtained by solving the problem

$$\min_d \ f(x+d) + \frac{\mu}{2}|d|^2 \tag{5.2}$$

lies on the fast track. This identification result is similar to the one we prove in Section 6.5, but the calculation of d is different. In our case of $f = h \circ c$, (5.2) becomes obtain

$$\min_d \ h\big(c(x+d)\big) + \frac{\mu}{2}|d|^2, \tag{5.3}$$

whose optimality conditions are, for some fixed current iterate x,

$$\nabla c(x+d)^* v + \mu d = 0, \ \text{ for some } v \in \partial h\big(c(x+d)\big). \tag{5.4}$$

Compare this system with the optimality conditions (4.2) from subproblem (4.1):

$$\nabla c(x)^* v + \mu d = 0, \ \text{ for some } v \in \partial h\big(c(x) + \nabla c(x) d\big).$$

In many applications of interest, c is nonlinear, so the subproblem (5.3) is generally harder to solve for the step d than our subproblem (4.1).

Mifflin and Sagastizábal [25] describe an algorithm in which an approximate solution of the subproblem (5.2) is obtained, again for the case of a convex objective, by making use of a piecewise linear underapproximation to their objective f. The approach is most suitable for a bundle method in which the piecewise-linear approximation is constructed from subgradients gathered at previous iterations. Approximations to the manifold of smoothness for f are constructed from the solution of this approximate proximal point calculation, and a Newton-like step for the Lagrangian is taken along this manifold, as envisioned in earlier methods. Daniilidis, Hare, and Malick [9] use the terminology "predictor-corrector" to describe algorithms of this type. Their "predictor" step is the step along the manifold of smoothness for f, while the "corrector" step (5.2) eventually returns the iterates to the correct active manifold (see [9, Theorem 28]). Miller and Malick [26] show how algorithms of this type are related to Newton-like methods that have been proposed earlier in various contexts.

Various of the algorithms discussed above make use of curvature information for the objective on the active manifold to accelerate local convergence. The algorithmic framework that we describe in Section 7 could easily be modified to incorporate similar techniques, while retaining its global convergence and manifold identification properties.

6. Properties of the Proximal Linearized Subproblem. We show in this section that when h is prox-regular at $\bar{c}$, under a mild additional assumption, the subproblem (4.1) has a local solution d with norm $O(|x - \bar{x}|)$, when the parameter μ is sufficiently large. When h is convex, this solution is the unique global solution of the subproblem. We show too that a point x_{new} near $x + d$ can be found such that the objective value $h\big(c(x_{\text{new}})\big)$ is close to the prediction of the linearized model $h(c(x) + \nabla c(x) d)$. Further, we describe conditions under which the subproblem correctly identifies the manifold $\mathcal{M}$ with respect to which h is partly smooth at the solution of (1.1).

6.1. Lipschitz Properties. We start with technical preliminaries. Allowing nonlipschitz or extended-valued outer functions h in our problem (1.1) is conceptually appealing, since it allows us to model constraints that must be enforced. However, this flexibility presents certain technical challenges, which we now address. We begin with a simple example, to illustrate some of the difficulties.

EXAMPLE 6.1. *Define a $\mathcal{C}^2$ function $c : \mathbb{R} \to \mathbb{R}^2$ by $c(x) = (x, x^2)$, and a lower semicontinuous convex function $h : \mathbb{R}^2 \to \bar{\mathbb{R}}$ by*

$$h(y,z) = \begin{cases} y & (z \geq 2y^2) \\ +\infty & (z < 2y^2). \end{cases}$$

The composite function $h \circ c$ is simply $\delta_{\{0\}}$, the indicator function of $\{0\}$. This function has a global minimum value zero, attained uniquely by $\bar{x} = 0$.

At any point $x \in \mathbb{R}$, the derivative map $\nabla c(x) : \mathbb{R} \to \mathbb{R}^2$ is given by $\nabla c(x)d = (d, 2xd)$ for $d \in \mathbb{R}$. Then, for all nonzero x, it is easy to check

$$h\big(c(x) + \nabla c(x)d\big) = +\infty \quad \text{for all } d \in \mathbb{R}^n,$$

so the corresponding proximal linearized subproblem (4.1) has no feasible solutions: its objective value is identically $+\infty$.

The adjoint map $\nabla c(\bar{x})^* : \mathbb{R}^2 \to \mathbb{R}$ is given by $\nabla c(\bar{x})^* v = v_1$ for $v \in \mathbb{R}^2$, and

$$\partial h(0,0) \;=\; \big\{v \in \mathbb{R}^2 : v_1 = 1,\ v_2 \leq 0\big\}.$$

Hence the criticality condition (1.3) has no solution $\bar{v} \in \mathbb{R}^2$.

This example illustrates two fundamental difficulties. The first is theoretical: the basic criticality condition (1.3) may be unsolvable, essentially because the chain rule fails. The second is computational: if, implicit in the function h, are constraints on acceptable values for $c(x)$, then curvature in these constraints can cause infeasibility in linearizations. As we see below, resolving both difficulties requires a kind of "transversality" condition common in variational analysis.

The transversality condition we need involves the "horizon subdifferential" of the function $h : \mathbb{R}^m \to \bar{\mathbb{R}}$ at the point $\bar{c} \in \mathbb{R}^m$, denoted $\partial^\infty h(\bar{c})$. This object, which recurs throughout our analysis, consists of a set of "horizon subgradients", capturing information about directions in which h grows faster than linearly near $\bar{c}$. Useful to keep in mind is the following fact:

$$\partial^\infty h(\bar{c}) = \{0\} \quad \text{if } h \text{ is locally Lipschitz around } \bar{c}.$$

This condition holds in particular for a convex function h that is continuous at $\bar{c}$. Readers interested only in continuous convex functions h may therefore make the substantial simplification $\partial^\infty h(\bar{c}) = \{0\}$ throughout the analysis. For general convex $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, for any point $\bar{c}$ in the domain $\operatorname{dom} h$ we have the following relationship between the horizon subdifferential and the classical normal cone to the domain (see [34, Proposition 8.12]):

$$\partial^\infty h(\bar{c}) = N_{\operatorname{dom} h}(\bar{c}).$$

We seek conditions guaranteeing a reasonable step in the proximal linearized subproblem (4.1). Our key tool is the following technical result.

THEOREM 6.1. *Consider a lower semicontinuous function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$, a point $\bar{z} \in \mathbb{R}^m$ where $h(\bar{z})$ is finite, and a linear map $\bar{G}\colon \mathbb{R}^n \to \mathbb{R}^m$ satisfying*

$$\partial^\infty h(\bar{z}) \cap \operatorname{Null}(\bar{G}^*) = \{0\}.$$

Then there exists a constant $\gamma > 0$ such that, for all vectors $z \in \mathbb{R}^n$ and linear maps $G : \mathbb{R}^n \to \mathbb{R}^m$ with (z, G) near $(\bar{z}, \bar{G})$, there exists a vector $w \in \mathbb{R}^m$ satisfying

$$|w| \leq \gamma |z - \bar{z}| \quad \text{and} \quad h(z + Gw) \leq h(\bar{z}) + \gamma |z - \bar{z}|.$$

Notice that this result is trivial if h is locally Lipschitz (or in particular continuous and convex) around $\bar{z}$, since we can simply choose $w = 0$. The nonlipschitz case is harder; our proof appears below following the introduction of a variety of ideas from variational analysis whose use is confined to this subsection. We refer the reader to Rockafellar and Wets [34] or Mordukhovich [27] for further details. First, we need a "metric regularity" result. Since this theorem is a fundamental tool for us, we give two proofs, one of which specializes the proof of Theorem 3.3 in Dontchev, Lewis and Rockafellar [11], while the other sets the result into a broader context.

THEOREM 6.2 (uniform metric regularity under perturbation). *Suppose that the closed set-valued mapping $F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ is metrically regular at a point $\bar{u} \in \mathbb{R}^p$ for a point $\bar{v} \in F(\bar{u})$: in other words, there exist constants $\kappa, a > 0$ such that all points $u \in B_a(\bar{u})$ and $v \in B_a(\bar{v})$ satisfy*

$$\operatorname{dist}\big(u, F^{-1}(v)\big) \leq \kappa \cdot \operatorname{dist}\big(v, F(u)\big). \tag{6.1}$$

Then there exist constants $\delta, \gamma > 0$ such that all linear maps $H\colon \mathbb{R}^p \to \mathbb{R}^q$ with $\|H\| < \delta$ and all points $u \in B_\delta(\bar{u})$ and $v \in B_\delta(\bar{v})$ satisfy

$$\operatorname{dist}\big(u, (F+H)^{-1}(v)\big) \;\leq\; \gamma \operatorname{dist}\big(v, (F+H)(u)\big). \tag{6.2}$$

Proof. For our first approach, we follow the notation of the proof of [11, Theorem 3.3]. Fix any constants

$$\lambda \in (0, \kappa^{-1}), \qquad \alpha \in \left(0, \frac{a}{4}(1 - \kappa\lambda)\min\{1, \kappa\}\right), \qquad \delta \in \left(0, \min\left\{\frac{\alpha}{4}, \frac{\alpha}{4\kappa}, \lambda\right\}\right).$$

Then the proof shows inequality (6.2), if we define $\gamma = \kappa/(1 - \kappa\lambda)$.

As an alternative, more formal approach, denote the space of linear maps from $\mathbb{R}^p$ to $\mathbb{R}^q$ by $L(\mathbb{R}^p, \mathbb{R}^q)$, and define a mapping $g\colon L(\mathbb{R}^p, \mathbb{R}^q) \times \mathbb{R}^p \to \mathbb{R}^q$ and a parametric mapping $g_H\colon \mathbb{R}^p \to \mathbb{R}^q$ by $g(H, u) = g_H(u) = Hu$ for maps $H \in L(\mathbb{R}^p, \mathbb{R}^q)$ and points $u \in \mathbb{R}^p$. Using the notation of [10, Section 3], the Lipschitz constant $l[g](0; \bar{u}, 0)$, is by definition the infimum of the constants ρ for which the inequality

$$d\big(v, g_H(u)\big) \leq \rho d\big(u, g_H^{-1}(v)\big)$$

holds for all triples (u, v, H) sufficiently near the triple $(\bar{u}, \bar{v}, 0)$. A quick calculation shows that this constant is zero. We can also consider $F + g$ as a set-valued mapping from $L(\mathbb{R}^p, \mathbb{R}^q) \times \mathbb{R}^p$ to $\mathbb{R}^q$, defined by $(F + g)(H, u) = F(u) + Hu$, and then the parametric mapping $(F + g)_H\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ is defined in the obvious way: in other words, $(F + g)_H(u) = F(u) + Hu$. According to [10, Theorem 2], we have the following relationship between the "covering rates" for F and $F + g$:

$$r[F + g](0; \bar{u}, \bar{v}) = r[F](\bar{u}, \bar{v}).$$

The reciprocal of the right-hand side is, by definition, the infimum of the constants $\kappa > 0$ such that inequality (6.1) holds for all pairs (u, v) sufficiently near the pair $(\bar{u}, \bar{v})$. By metric regularity, this number is strictly positive. On the other hand,

the reciprocal of the left-hand side is, by definition, the infimum of the constants $\gamma > 0$ such that inequality (6.2) holds for all triples (u, v, H) sufficiently near the pair $(\bar{u}, \bar{v}, 0)$. □

The following result depends on an assumption about the *normal cone* to S at a point $s \in S$, denoted $N_S(s)$, the basic building block for variational analysis (see Rockafellar and Wets [34] or Mordukhovich [27]). When S is convex, it coincides exactly with the classic normal cone from convex analysis, while for smooth manifolds it coincides with the classical normal space.

COROLLARY 6.3. *Consider a closed set $S \subset \mathbb{R}^q$ with $0 \in S$, and a linear map $\bar{A}\colon \mathbb{R}^p \to \mathbb{R}^q$ satisfying*

$$N_S(0) \cap \mathrm{Null}(\bar{A}^*) = \{0\}.$$

Then there exists a constant $\gamma > 0$ such that, for all vectors $v \in \mathbb{R}^q$ and linear maps $A : \mathbb{R}^p \to \mathbb{R}^q$ with (v, A) near $(0, \bar{A})$, the inclusion

$$v + Au \in S$$

has a solution $u \in \mathbb{R}^p$ satisfying $|u| \leq \gamma|v|$.

Proof. Corresponding to any linear map $A\colon \mathbb{R}^p \to \mathbb{R}^q$, define a set-valued mapping $F_A\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ by $F_A(u) = Au - S$. A coderivative calculation shows, for vectors $v \in \mathbb{R}^p$,

$$D^*F_A(0|0)(v) = \begin{cases} \{A^*v\} & \big(v \in N_S(0)\big) \\ \emptyset & (\text{otherwise}). \end{cases}$$

Hence, by assumption, the only vector $v \in \mathbb{R}^p$ satisfying $0 \in D^*F_{\bar{A}}(0|0)(v)$ is zero, so by [34, Thm 9.43], the mapping $F_{\bar{A}}$ is metrically regular at zero for zero. Applying the preceding theorem shows that there exist constants $\delta, \gamma > 0$ such that, if $\|A - \bar{A}\| < \delta$ and $|v| < \delta$, then we have

$$\mathrm{dist}\big(0, F_A^{-1}(-v)\big) \leq \gamma \mathrm{dist}\big(-v, F_A(0)\big),$$

or equivalently,

$$\mathrm{dist}\big(0, A^{-1}(S - v)\big) \leq \gamma \mathrm{dist}\,(v, S).$$

Since $0 \in S$, the right-hand side is bounded above by $\gamma|v|$, so the result follows. □

We are now ready to prove the result we claimed at the outset of this subsection.

Proof of Theorem 6.1. Without loss of generality, we can suppose $\bar{z} = 0$ and $h(0) = 0$. Let $S \subset \mathbb{R}^m \times \mathbb{R}$ be the epigraph of h, and define a map $\bar{A}\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m \times \mathbb{R}$ by $\bar{A}(z, \tau) = (\bar{G}z, \tau)$. Clearly we have $\mathrm{Null}(\bar{A}^*) = \mathrm{Null}(\bar{G}^*) \times \{0\}$, so [34, Theorem 8.9] shows

$$N_S(0, 0) \cap \mathrm{Null}(\bar{A}^*) = \{(0, 0)\}.$$

For any vector z and linear map G with (z, G) near $(\bar{z}, \bar{G})$, the vector $(z, 0) \in \mathbb{R}^m \times \mathbb{R}$ is near the vector $(\bar{z}, 0)$ and the map $(w, \tau) \mapsto (Gw, \tau)$ is near the map $(w, \tau) \mapsto (\bar{G}w, \tau)$. The previous corollary shows the existence of a constant $\gamma > 0$ such that, for all such z and G, the inclusion

$$(z, 0) + (Gw, \tau) \in S$$

has a solution satisfying $|(w,\tau)| \le \gamma|(z,0)|$, and the result follows. $\square$

We end this subsection with another tool to be used later: the proof is a straightforward application of standard ideas from variational analysis. Like Theorem 6.2, this tool concerns metric regularity, this time for a constraint system of the form $F(z) \in S$ for an unknown vector z, where the map F is smooth, and S is a closed set.

THEOREM 6.4 (metric regularity of constraint systems). *Consider a $\mathcal{C}^1$ map $F\colon \mathbb{R}^p \to \mathbb{R}^q$, a point $\bar{z} \in \mathbb{R}^p$, and a closed set $S \subset \mathbb{R}^q$ containing the vector $F(\bar{z})$. Suppose the condition*

$$N_S\big(F(\bar{z})\big) \cap \operatorname{Null}(\nabla F(\bar{z})^*) \;=\; \{0\}$$

holds. Then there exists a constant $\kappa > 0$ such that all points $z \in \mathbb{R}^p$ near $\bar{z}$ satisfy the inequality

$$\operatorname{dist}\big(z, F^{-1}(S)\big) \le \kappa \cdot \operatorname{dist}(F(z), S).$$

Proof. We simply need to check that the set-valued mapping $G\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ defined by $G(z) = F(z) - S$ is metrically regular $\bar{z}$ for zero. Much the same coderivative calculation as in the proof of Corollary 6.3 shows, for vectors $v \in \mathbb{R}^p$, the formula

$$D^*G(\bar{z}|0)(v) = \begin{cases} \{\nabla F(\bar{z})^* v\} & \big(v \in N_S(\bar{z})\big) \\ \emptyset & (\text{otherwise}). \end{cases}$$

Hence, by assumption, the only vector $v \in \mathbb{R}^p$ satisfying $0 \in D^*G(\bar{z}|0)(v)$ is zero, so metric regularity follows by [34, Thm 9.43]. $\square$

6.2. The Proximal Step. We now prove a key result. Under a standard transversality condition, and assuming the proximal parameter μ is sufficiently large (if the function h is nonconvex), we show the existence of a step $d = O(|x - \bar{x}|)$ in the proximal linearized subproblem (4.1) with corresponding objective value close to the critical value $h(\bar{c})$.

When the outer function h is locally Lipschitz (or, in particular, continuous and convex), this result and its proof simplify considerably. First, the transversality condition is automatic. Second, while the proof of the result appeals to the technical tool we developed in the previous subsection (Theorem 6.1), this tool is trivial in the Lipschitz case, as we noted earlier.

THEOREM 6.5 (proximal step). *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$. Assume the transversality condition*

$$\partial^\infty h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*) \;=\; \{0\}. \tag{6.3}$$

Then there exist numbers $\bar{\mu} \ge 0$, $\delta > 0$, and $\rho \ge 0$, and a mapping $d : B_\delta(\bar{x}) \times (\bar{\mu}, \infty) \to \mathbb{R}^n$ such that the following properties hold.

- *(a) For all points $x \in B_\delta(\bar{x})$ and all parameter values $\mu > \bar{\mu}$, the step $d(x,\mu)$ is a local minimizer of the proximal linearized subproblem (4.1), and moreover $|d(x,\mu)| \le \rho|x - \bar{x}|$.*
- *(b) Given any sequences $x_r \to \bar{x}$ and $\mu_r > \bar{\mu}$, then if either $\mu_r|x_r - \bar{x}|^2 \to 0$ or $h\big(c(x_r)\big) \to h(\bar{c})$, we have*

$$h\big(c(x_r) + \nabla c(x_r) d(x_r, \mu_r)\big) \to h(\bar{c}). \tag{6.4}$$

(c) When h is convex and lower semicontinuous, the results of parts (a) and (b) hold with $\bar{\mu} = 0$.

Proof. Without loss of generality, suppose $\bar{x} = 0$ and $\bar{c} = c(0) = 0$, and furthermore $h(0) = 0$. By assumption,

$$0 \in \partial(h \circ c)(0) \subset \nabla c(0)^* \partial h(0),$$

using the chain rule [34, Thm 10.6], so there exists a vector

$$v \in \partial h(0) \cap \mathrm{Null}(\nabla c(0)^*).$$

We first prove part (a). By prox-regularity, there exists a constant $\rho \geq 0$ such that

$$h(z) \geq \langle v, z \rangle - \frac{\rho}{2}|z|^2 \tag{6.5}$$

for all small vectors $z \in \mathbb{R}^m$. Hence, there exists a constant $\delta_1 > 0$ such that ∇c is continuous on $B_{\delta_1}(0)$ and

$$h_{x,\mu}(d) \;\geq\; \langle v, c(x) + \nabla c(x)d \rangle - \frac{\rho}{2}|c(x) + \nabla c(x)d|^2 + \frac{\mu}{2}|d|^2$$

for all vectors $x, d \in B_{\delta_1}(0)$. As a consequence, we have that

$$h_{x,\mu}(d) \;\geq\; \min_{|x| \leq \delta_1,\, |d| = \delta_1} \left\{ \langle v, c(x) + \nabla c(x)d \rangle - \frac{\rho}{2}|c(x) + \nabla c(x)d|^2 \right\} + \frac{\mu}{2}|d|^2,$$

and the term in braces is finite by continuity of c and ∇c on $B_{\delta_1}(0)$. Hence by choosing $\bar{\mu}$ sufficiently large (certainly greater than $\rho\|\nabla c(0)\|^2$) we can ensure that

$$h_{x,\bar{\mu}}(d) \geq 1 \quad \text{whenever } |x| \leq \delta_1,\ |d| = \delta_1.$$

Then for $x \in B_{\delta_1}(0)$, $|d| = \delta_1$, and $\mu \geq \bar{\mu}$, we have

$$h_{x,\mu}(d) = h_{x,\bar{\mu}}(d) + \frac{1}{2}(\mu - \bar{\mu})|d|^2 \geq 1 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2. \tag{6.6}$$

Since c is $\mathcal{C}^2$ at 0, there exist constants $\beta > 0$ and $\delta_2 \in (0, \delta_1)$ such that, for all $x \in B_{\delta_2}(0)$, the vector

$$z(x) = c(x) - \nabla c(x)x \tag{6.7}$$

satisfies $|z(x)| \leq \beta|x|^2$. Setting $G = \nabla c(x)$, $\bar{G} = \nabla c(0)$, $\bar{z} = 0$, and $z = z(x)$, we now apply Theorem 6.1. Hence for some constants $\gamma > 0$ and $\delta_3 \in (0, \delta_2)$, given any vector $x \in B_{\delta_3}(0)$, there exists a vector $d \in \mathbb{R}^n$ (defined by $d = w - x$, in the notation of the theorem) satisfying

$$\begin{aligned} |x + d| &\leq \gamma|z(x)| \leq \gamma\beta|x|^2 \\ h\big(c(x) + \nabla c(x)d\big) &\leq \gamma|z(x)| \leq \gamma\beta|x|^2. \end{aligned}$$

We deduce the existence of a constant $\delta_4 \in (0, \delta_3)$ such that, for all $x \in B_{\delta_4}(0)$, the corresponding d satisfies

$$|d| \leq |x| + \gamma\beta|x|^2 \;<\; \delta_1,$$

and

$$\begin{aligned} h_{x,\mu}(d) &= h(c(x) + \nabla c(x)d) + \frac{\mu}{2}|d|^2 \\ &\leq \gamma\beta|x|^2 + \frac{\bar{\mu}}{2}\big(|x| + \gamma\beta|x|^2\big)^2 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2 \\ &< 1 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2. \end{aligned}$$

We denote this d by $\hat{d}(x)$.

The lower semicontinuous function $h_{x,\mu}$ must have a minimizer (which we denote $d(x,\mu)$) over the compact set $B_{\delta_1}(0)$, and the inequality above implies the corresponding minimum value is majorized by $h_{x,\mu}\big(\hat{d}(x)\big)$, and thus is strictly less than $1 + (1/2)(\mu - \bar{\mu})\delta_1^2$. But inequality (6.6) implies that this minimizer must lie in the interior of the ball $B_{\delta_1}(0)$; in particular, it must be an unconstrained local minimizer of $h_{x,\mu}$. By setting $\delta = \delta_4$, we complete the proof of the first part of (a). Notice furthermore that for $x \in B_{\delta_4}(0)$, we have

$$\begin{aligned} &h\big(c(x) + \nabla c(x)d(x,\mu)\big) \\ &\qquad \leq\ h_{x,\mu}\big(d(x,\mu)\big)\ \leq\ h_{x,\mu}\big(\hat{d}(x)\big)\ \leq\ \gamma\beta|x|^2 + \frac{\mu}{2}\big(|x| + \gamma\beta|x|^2\big)^2. \end{aligned} \tag{6.8}$$

We now prove the remainder of part (a), that is, uniform boundedness of the ratio $|d(x,\mu)|/|x|$. Suppose there are sequences $x_r \in B_\delta(\bar{x})$ and $\mu_r \geq \bar{\mu}$ such that $|d(x_r,\mu_r)|/|x_r| \to \infty$. Since $|d(x_r,\mu_r)| \leq \delta_1$ by the arguments above, we must have $x_r \to 0$. By the arguments above, for all large r we have the following inequalities:

$$\begin{aligned} &\gamma\beta|x_r|^2 + \frac{\mu_r}{2}\big(|x_r| + \gamma\beta|x_r|^2\big)^2 \\ &\quad \geq h_{x_r,\mu_r}(d_r) \\ &\quad \geq \langle v, c(x_r) + \nabla c(x_r)d_r\rangle - \frac{\rho}{2}|c(x_r) + \nabla c(x_r)d_r|^2 + \frac{\mu_r}{2}|d_r|^2. \end{aligned}$$

Dividing each side by $(1/2)\mu_r|x_r|^2$ and letting $r \to \infty$, we recall

$$\mu_r \geq \bar{\mu} > \rho\|\nabla c(0)\|^2 \geq 0$$

and observe that the left-hand side remains finite, while the right-hand side is eventually dominated by $(1 - \rho\|\nabla c(0)\|^2/\mu_r)|d_r|^2/|x_r|^2$, which approaches ∞, yielding a contradiction.

For part (b), suppose first that $\mu_r|x_r|^2 \to 0$. By substituting $(x,\mu) = (x_r,\mu_r)$ into (6.8), we have that

$$\limsup\ h\big(c(x_r) + \nabla c(x_r)d(x_r,\mu_r)\big)\ \leq\ 0. \tag{6.9}$$

From part (a), we have that $|d(x_r,\mu_r)|/|x_r|$ is uniformly bounded, hence $d(x_r,\mu_r) \to 0$ and thus $c(x_r) + \nabla c(x_r)d(x_r,\mu_r) \to 0$. Being prox-regular, h is lower semicontinuous at 0, so

$$\liminf\ h\big(c(x_r) + \nabla c(x_r)d(x_r,\mu_r)\big)\ \geq\ 0.$$

By combining these last two inequalities, we obtain

$$h\big(c(x_r) + \nabla c(x_r)d(x_r,\mu_r)\big) \to 0,$$

as required.

Now suppose instead that $h\big(c(x_r)\big) \to h(\bar{c}) = 0$. We have from (6.8) that

$$h\big(c(x_r) + \nabla c(x_r) d(x_r, \mu_r)\big) \le h_{x_r,\mu_r}\big(d(x_r,\mu_r)\big) \le h_{x_r,\mu_r}(0) = h\big(c(x_r)\big).$$

By taking the lim sup of both sides we again obtain (6.9), and the result follows as before.

For part (c), when h is lower semicontinuous and convex, the argument simplifies. We set $\rho = 0$ in (6.5) and choose the constant $\delta > 0$ so the map ∇c is continuous on $B_\delta(0)$. Choosing the constants β and γ as before, Theorem 6.1 again guarantees the existence, for all small points x, of a step $\hat{d}(x)$ satisfying

$$h\big(c(x) + \nabla c(x)\hat{d}(x)\big) \le \gamma\beta|x|^2.$$

We now deduce that the proximal linearized objective $h_{x,\mu}$ is somewhere finite, so has compact level sets, by coercivity. Thus it has a global minimizer $d(x,\mu)$ (unique, by strict convexity), which must satisfy the inequality

$$h\big(c(x) + \nabla c(x) d(x,\mu)\big) \le h\big(c(x) + \nabla c(x)\hat{d}(x)\big) \le \gamma\beta|x|^2.$$

The remainder of the argument proceeds as before. □

We discuss Theorem 6.5(b) by giving a simple example of a function prox-regular at $c(\bar{x})$ such that for sequences $x_r \to \bar{x}$ and $\mu_r \to \infty$ that satisfy neither $\mu_r|x_r - \bar{x}|^2 \to 0$ nor $h\big(c(_r)\big) \to h\big(c(\bar{x})\big)$, the conclusion (6.4) fails to hold. For a scalar x, take $c(x) = x$ and

$$h(c) = \begin{cases} -c & (c \le 0) \\ 1 + c & (c > 0). \end{cases}$$

The unique critical point is clearly $\bar{x} = 0$ with $c(\bar{x}) = 0$ and $h\big(c(\bar{x})\big) = 0$, and this problem satisfies the assumptions of the theorem. Consider $x > 0$, for which the subproblem (4.1) is

$$\min_d h_{x,\mu}(d) = h(x+d) + \frac{\mu}{2}d^2 = \begin{cases} -x - d + \frac{\mu}{2}d^2 & (x + d \le 0) \\ 1 + x + d + \frac{\mu}{2}d^2 & (x + d > 0). \end{cases}$$

When $\mu_r x_r \in (0, 1]$, then $d_r = -x_r$ is the only local minimizer of h_{x_r,μ_r}. When $\mu_r x_r > 1$, the situation is more interesting. The value $d_r = -\mu_r^{-1}$ minimizes the "positive" branch of h_{x_r,μ_r}, with function value $1 + x_r - (2\mu_r)^{-1}$, and there is a second local minimizer at $d_r = -x_r$, with function value $(\mu_r/2)x_r^2$. (In both cases, the local minimizers satisfy the estimate $|d_r| = O(|x_r - \bar{x}|)$ proved in part (a).) Comparison of the function values show that in fact the global minimum is achieved at the former point—$d_r = -\mu_r^{-1}$—when $x_r > \mu_r^{-1} + \sqrt{2}\mu_r^{-1/2}$. If this step is taken, we have $x_r + d_r > 0$, so the new iterate remains on the upper branch of h. For sequences $x_r = \mu_r^{-1} + 2\mu_r^{-1/2}$ and $\mu_r \to \infty$, we thus have for the global minimizer $d_r = -\mu_r$ of h_{x_r,μ_r} that $h(c(x_r) + \nabla c(x_r) d_r) > 1$ for all r, while $h\big(c(\bar{x})\big) = 0$, so that (6.4) does not hold.

6.3. Restoring Feasibility. In the algorithmic framework that we have in mind, the basic iteration starts at a current point $x \in \mathbb{R}^n$ such that the function h is finite at the vector $c(x)$. We then solve the proximal linearized subproblem (4.1) to obtain the

step $d = d(x,\mu) \in \mathbb{R}^n$ discussed earlier in this section. Under reasonable conditions we showed that, for x near the critical point $\bar{x}$, we have $d = O(|x - \bar{x}|)$ and furthermore we know that the value of h at the vector $c(x) + \nabla c(x)d$ is close to the critical value $h\big(c(\bar{x})\big)$.

The algorithmic idea is now to update the point x to a new point $x + d$. When the function h is Lipschitz, this update is motivated by the fact that, since the map c is $\mathcal{C}^2$, we have, uniformly for x near the critical point $\bar{x}$,

$$c(x+d) - \big(c(x) + \nabla c(x)d\big) \;=\; O(|d|^2)$$

and hence

$$h\big(c(x+d)\big) - h\big((c(x) + \nabla c(x)d)\big) \;=\; O(|d|^2).$$

However, if h is not Lipschitz, it may not be appropriate to update x to $x + d$: the value $h\big(c(x+d)\big)$ may even be infinite.

In order to take another basic iteration, we need somehow to restore the point $x + d$ to feasibility, or more generally to find a nearby point with objective value not much worse than our linearized estimate $h(c(x) + \nabla c(x)d)$. Depending on the form of the function h, this may or may not be easy computationally. However, as we now discuss, our fundamental transversality condition (6.3), guarantees that such a restoration is always possible in theory. In the next section, we refer to this restoration process as an "efficient projection."

THEOREM 6.6 (linear estimator improvement). *Consider a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$ that is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$, and a lower semicontinuous function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ that is finite at the vector $\bar{c} = c(\bar{x})$. Assume the transversality condition (6.3) holds, namely*

$$\partial^\infty h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*) \;=\; \{0\}.$$

Then there exists constants $\gamma, \delta > 0$ such that, for any point $x \in B_\delta(\bar{x})$ and any step $d \in B_\delta(0) \subset \mathbb{R}^n$ for which $|h(c(x)+\nabla c(x)d) - h(\bar{c})| < \delta$, there exists a point $x_{\rm new} \in \mathbb{R}^n$ satisfying

$$|x_{\rm new} - (x+d)| \le \gamma|d|^2 \quad \textit{and} \quad h\big(c(x_{\rm new})\big) \;\le\; h(c(x) + \nabla c(x)d) + \gamma|d|^2. \tag{6.10}$$

Proof. Define a $\mathcal{C}^2$ map $F\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m \times \mathbb{R}$ by $F(x,t) = (c(x), t)$. Notice that the epigraph $\operatorname{epi} h$ is a closed set containing the vector $F\big(\bar{x}, h(\bar{c})\big)$. Clearly we have

$$\operatorname{Null}\Big(\nabla F\big(\bar{x}, h(\bar{c})\big)^*\Big) \;=\; \operatorname{Null}\big(\nabla c(\bar{x})^*\big) \times \{0\}.$$

On the other hand, using Rockafellar and Wets [34, Theorem 8.9], we have

$$(y,0) \in N_{\operatorname{epi} h}\big(\bar{c}, h(\bar{c})\big) \quad \Leftrightarrow \quad y \in \partial^\infty h(\bar{c}).$$

Hence the transversality condition is equivalent to

$$N_{\operatorname{epi} h}\big(\bar{c}, h(\bar{c})\big) \cap \operatorname{Null}\Big(\nabla F\big(\bar{x}, h(\bar{c})\big)^*\Big) \;=\; \{0\}.$$

We next apply Theorem 6.4 to deduce the existence of a constant $\kappa > 0$ such that, for all vectors (u,t) near the vector $\big(\bar{x}, h(\bar{c})\big)$ we have

$$\operatorname{dist}\big((u,t), F^{-1}(\operatorname{epi} h)\big) \;\le\; \kappa \cdot \operatorname{dist}(F(u,t), \operatorname{epi} h).$$

Thus there exists a constant $\delta > 0$ such that, for any point $x \in B_\delta(\bar{x})$ and any step $d \in \mathbb{R}^n$ satisfying $|d| \leq \delta$ and $|h(c(x) + \nabla c(x)d) - h(\bar{c})| < \delta$, we have

$$\begin{aligned}
&\operatorname{dist}\Big(\big(x+d, h(c(x)+\nabla c(x)d)\big), F^{-1}(\operatorname{epi} h)\Big) \\
&\qquad \leq \ \kappa \cdot \operatorname{dist}\big(F\big(x+d, h(c(x)+\nabla c(x)d)\big), \operatorname{epi} h\big) \\
&\qquad = \ \kappa \cdot \operatorname{dist}\Big(\big(c(x+d), h(c(x)+\nabla c(x)d)\big), \operatorname{epi} h\Big) \\
&\qquad \leq \ \kappa \cdot |c(x+d) - (c(x)+\nabla c(x)d)|,
\end{aligned}$$

since

$$\big(c(x)+\nabla c(x)d, h(c(x)+\nabla c(x)d)\big) \ \in \ \operatorname{epi} h.$$

Since the map c is $\mathcal{C}^2$, by reducing δ if necessary we can ensure the existence of a constant $\gamma > 0$ such that the right-hand side of the above chain of inequalities is bounded above by $\gamma|d|^2$.

We have therefore shown the existence of a vector

$$(x_{\text{new}}, t) \in F^{-1}(\operatorname{epi} h)$$

satisfying the inequalities

$$|x_{\text{new}} - (x+d)| \leq \gamma|d|^2 \ \text{ and } \ |t - h(c(x)+\nabla c(x)d)| \ \leq \ \gamma|d|^2.$$

We therefore know $t \geq h\big(c(x_{\text{new}})\big)$, so the result follows. □

6.4. Uniqueness of the Proximal Step and Convergence of Multipliers. Our focus in this subsection is on uniqueness of the local solution of (4.1) near 0, uniqueness of the corresponding multiplier vector, and on showing that the solution $d(x,\mu)$ of (4.1) has a strictly lower subproblem objective value than $d = 0$. For the uniqueness results, we strengthen the transversality condition (6.3) to a constraint qualification that we now introduce.

Throughout this subsection we assume that the function h is prox-regular at the point $\bar{c}$. Since prox-regular functions are (Clarke) subdifferentially regular, the subdifferential $\partial h(\bar{c})$ is a closed and convex set in $\mathbb{R}^m$, and its recession cone is exactly the horizon subdifferential $\partial^\infty h(\bar{c})$ (see [34, Corollary 8.11]). Denoting the subspace parallel to the affine span of the subdifferential by $\operatorname{par} \partial h(\bar{c})$, we deduce that

$$\partial^\infty h(\bar{c}) \subset \operatorname{par} \partial h(\bar{c}).$$

Hence the "constraint qualification" that we next consider, namely

$$\operatorname{par} \partial h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*) \ = \ \{0\} \tag{6.11}$$

implies the transversality condition (6.3).

Condition (6.11) is related to the linear independence constraint qualification in classical nonlinear programming. To illustrate, consider again the case of Section 3, where the function h is finite and polyhedral:

$$h(c) = \max_{i \in I}\{\langle h_i, c\rangle + \beta_i\}$$

for some given vectors $h_i \in \mathbb{R}^m$ and scalars β_i. Then, as we noted,

$$\partial h(\bar{c}) = \operatorname{conv}\{h_i : i \in \bar{I}\},$$

where $\bar{I}$ is the set of active indices, so

$$\operatorname{par} \partial h(\bar{c}) \;=\; \Big\{ \sum_{i\in\bar{I}} \lambda_i h_i : \sum_{i\in\bar{I}} \lambda_i = 0 \Big\}.$$

Thus condition (6.11) states

$$\sum_{i\in\bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \sum_{i\in\bar{I}} \lambda_i h_i = 0. \tag{6.12}$$

By contrast, the linear independence constraint qualification for the corresponding nonlinear program (3.3) at the point $\big(\bar{x}, -h(\bar{c})\big)$ is

$$\sum_{i\in\bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \lambda_i = 0 \;\; (i \in \bar{I}),$$

which is a stronger assumption than condition (6.12).

We now prove a straightforward technical result that addresses two issues: existence and boundedness of multipliers for the proximal subproblem (4.1), and the convergence of these multipliers to a unique multiplier that satisfies criticality conditions for (1.1), when the constraint qualification (6.11) is satisfied. The argument is routine but, as usual, it simplifies considerably in the case of h locally Lipschitz (or in particular convex and continuous) around the point $\bar{c}$, since then the horizon subdifferential $\partial^\infty h$ is identically $\{0\}$ near $\bar{c}$.

LEMMA 6.7. *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$.*

When the transversality condition (6.3) holds, then for any sequences $\mu_r > 0$ and $x_r \to \bar{x}$ such that $\mu_r |x_r - \bar{x}| \to 0$, and any sequence of critical points $d_r \in \mathbb{R}^n$ for the corresponding proximal linearized subproblems (4.1) satisfying the conditions

$$d_r = O(|x_r - \bar{x}|) \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r) d_r\big) \to h(\bar{c}),$$

there exists a bounded sequence of vectors $v_r \in \mathbb{R}^m$ that satisfy

$$0 = \nabla c(x_r)^* v_r + \mu_r d_r, \tag{6.13a}$$

$$v_r \in \partial h\big(c(x_r) + \nabla c(x_r) d_r\big). \tag{6.13b}$$

When the stronger constraint qualification (6.11) holds, in place of (6.3), the set of multipliers $v \in \mathbb{R}^m$ solving the criticality condition (1.3), namely

$$\partial h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*) \tag{6.14}$$

is in fact a singleton $\{\bar{v}\}$. Furthermore, any sequence of multipliers $\{v_r\}$ satisfying the conditions above converges to $\bar{v}$.

Proof. We first assume (6.3), and claim that

$$\partial^\infty h\big(c(x_r) + \nabla c(x_r) d_r\big) \cap \operatorname{Null}(\nabla c(x_r)^*) \;=\; \{0\} \tag{6.15}$$

for all large r. Indeed, if this property should fail, then for infinitely many r there would exist a unit vector v_r lying in the intersection on the left-hand side, and any limit point of these unit vectors must lie in the set

$$\partial^\infty h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*), \tag{6.16}$$

by outer semicontinuity of the set-valued mapping $\partial^\infty h$ at the point $\bar{c}$ [34, Proposition 8.7], contradicting the transversality condition (6.3). As a consequence, we can apply the chain rule to deduce the existence of vectors $v_r \in \mathbb{R}^n$ satisfying (6.13). This sequence must be bounded, since otherwise, after taking a subsequence, we could suppose $|v_r| \to \infty$ and then any limit point of the unit vectors $|v_r|^{-1}v_r$ would lie in the set (6.16), again contradicting the transversality condition. The first claim of the theorem is proved.

For the second claim, we assume the constraint qualification (6.11) and note as above that it implies the transversality condition (6.3), so the chain rule implies that the set (6.14) is nonempty. This set must therefore be a singleton $\{\bar{v}\}$, using (6.11) again. Using boundedness of $\{v_r\}$, and the fact that $\mu_r d_r \to 0$, we have by taking limits in (6.13) that any limit point of $\{v_r\}$ lies in (6.14) (by outer semicontinuity of ∂h at $\bar{c}$), and therefore $v_r \to \bar{v}$. □

Using Theorem 6.5, we show that the local minimizers of h_{x_r,μ_r} satisfy the desired properties, and in addition give a strict improvement over 0 in the subproblem (4.1).

LEMMA 6.8. *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, that the composite function $h \circ c$ is critical at $\bar{x}$, and that the transversality condition (6.3) holds. Defining $\bar{\mu}$ as in Theorem 6.5, let $\mu_r > \bar{\mu}$ and $x_r \to \bar{x}$ be sequences such that $\mu_r|x_r - \bar{x}| \to 0$. Then for all r sufficiently large, there is a local minimizer d_r of h_{x_r,μ_r} such that*

$$d_r = O(|x_r - \bar{x}|) \quad \text{and} \quad h(c(x_r) + \nabla c(x_r)d_r) \to h(\bar{c}). \tag{6.17}$$

Moreover, if $0 \notin \partial(h \circ c)(x_r)$ for all r, then $d_r \neq 0$ and

$$h_{x_r,\mu_r}(d_r) < h_{x_r,\mu_r}(0) \tag{6.18}$$

for all r sufficiently large.

Proof. Existence of a sequence of local minimizers d_r of h_{x_r,μ_r} with the properties (6.17) follows from parts (a) and (b) of Theorem 6.5 when we set $d_r = d(x_r, \mu_r)$ and use $\mu_r > \bar{\mu}$. We now prove (6.18). From (6.17) and Lemma 6.7 we deduce the existence of v_r satisfying (6.13). If we were to have $d_r = 0$, then these conditions reduce to

$$\nabla c(x_r)^* v_r = 0, \qquad v_r \in \partial h\big(c(x_r)\big),$$

so that $0 \in \partial(h \circ c)(x_r)$, by subdifferential regularity of h. Hence we must have $d_r \neq 0$.

By prox-regularity, we have

$$\begin{aligned}
h\big(c(x_r)\big) &\geq h(c(x_r) + \nabla c(x_r)d_r) + \langle v_r, -\nabla c(x_r)d_r\rangle - \frac{\rho}{2}|\nabla c(x_r)d_r|^2 \\
&= h(c(x_r) + \nabla c(x_r)d_r) + \mu_r|d_r|^2 - \frac{\rho}{2}|\nabla c(x_r)d_r|^2 \qquad \text{by (6.13a)} \\
&= h(c(x_r) + \nabla c(x_r)d_r) + \frac{\mu_r}{2}|d_r|^2 + \frac{\mu_r - \rho\|c(x_r)\|^2}{2}|d_r|^2 \\
&= h_{x_r,\mu_r}(d_r) + \frac{\mu_r - \rho\|c(x_r)\|^2}{2}|d_r|^2 \qquad \text{by (4.1)} \\
&> h_{x_r,\mu_r}(d_r),
\end{aligned}$$

where the final inequality holds because $\bar{\mu} > \rho\|\nabla c(\bar{x})\|^2$. □

Returning to the assumptions of Theorem 6.5, but now with the constraint qualification (6.11) replacing the weaker transversality condition (6.3), we can derive local uniqueness results about critical points for the proximal linearized subproblem.

When the outer function h is convex, uniqueness is obvious, since then the proximal linearized objective $h_{\mu,x}$ is strictly convex for any $\mu > 0$. For lower $\mathcal{C}^2$ functions, the argument is much the same: such functions have the form $g - \kappa|\cdot|^2$, locally, for some continuous convex function g, so again $h_{\mu,x}$ is locally strictly convex for large μ. For general prox-regular functions, the argument requires slightly more care.

THEOREM 6.9 (unique step). *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$. Suppose furthermore that the constraint qualification (6.11) holds. Then there exists $\hat{\mu} \geq 0$ such that the following properties hold. Given any sequence $\{\mu_r\}$ with $\mu_r > \hat{\mu}$ for all r and any sequence $x_r \to \bar{x}$ such that $\mu_r|x_r - \bar{x}| \to 0$, there exists a sequence of local minimizers d_r of h_{x_r,μ_r} and a corresponding sequence of multipliers v_r with the following properties:*

$$0 \in \partial h_{x_r,\mu_r}(d_r), \quad d_r = O(|x_r - \bar{x}|), \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r)d_r\big) \to h(\bar{c}), \tag{6.19}$$

as $r \to \infty$, and satisfying (6.13), with $v_r \to \bar{v}$, where $\bar{v}$ is the unique vector that solves the criticality condition (1.3). Moreover, d_r is uniquely defined for all r sufficiently large.

In the case of a convex, lower semicontinuous function $h : \mathbb{R}^m \to (-\infty, +\infty]$, the result holds with $\hat{\mu} = 0$.

Proof. The existence of sequences $\{d_r\}$ and $\{v_r\}$ with the claimed properties follows from Theorem 6.5 and Lemma 6.7. We need only prove the claim about uniqueness of the vectors d_r, and the final claim about the special case of h convex and lower semicontinuous.

Throughout the proof, we choose $\hat{\mu} > \bar{\mu}$, where $\bar{\mu}$ is defined in Theorem 6.5.

We first show the uniqueness of d_r in the general case. Since the function h is prox-regular at $c(\bar{x})$, its subdifferential ∂h has a *hypomonotone localization* around the point $(c(\bar{x}), \bar{v})$. In other words, there exist constants $\rho > 0$ and $\epsilon > 0$ such that the mapping $T\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ defined by

$$T(y) = \begin{cases} \partial h(y) \cap B_\epsilon(\bar{v}) & (y \in B_\epsilon\big(c(\bar{x})\big),\ |h(y) - h\big(c(\bar{x})\big)| \leq \epsilon) \\ \emptyset & (\text{otherwise}) \end{cases}$$

has the property

$$z \in T(y) \text{ and } z' \in T(y') \quad \Rightarrow \quad \langle z' - z, y' - y\rangle \geq -\rho|y' - y|^2.$$

(See [34, Example 12.28 and Theorem 13.36].) If the uniqueness claim does not hold, we have by taking a subsequence if necessary that there is a sequence $x_r \to \bar{x}$ and distinct sequences of $d_r^1 \neq d_r^2$ in $\mathbb{R}^n$ satisfying the conditions

$$0 \in \partial h_{x_r,\mu_r}(d_r^i), \quad d_r^i = O(|x_r - \bar{x}|) \to 0, \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r)d_r^i\big) \to h\big(c(\bar{x})\big),$$

as $r \to \infty$, for $i = 1, 2$. Lemma 6.7 shows the existence of sequences of vectors $v_r^i \in \mathbb{R}^n$ satisfying

$$\begin{aligned} 0 &= \nabla c(x_r)^* v_r^i + \mu_r d_r^i \\ v_r^i &\in \partial h\big(c(x_r) + \nabla c(x_r)d_r^i\big), \end{aligned}$$

for all large r, and furthermore $v_r^i \to \bar{v}$ for each $i = 1, 2$. Consequently, for all large r we have

$$v_r^i \in T\big(c(x_r) + \nabla c(x_r)d_r^i\big) \quad \text{for } i = 1, 2,$$

so that

$$-\mu_r|d_r^1 - d_r^2|^2 \;=\; \langle v_r^1 - v_r^2, \nabla c(x_r)(d_r^1 - d_r^2)\rangle \;\geq\; -\rho|\nabla c(x_r)(d_r^1 - d_r^2)|^2.$$

Since $\hat{\mu} > \bar{\mu} > \rho\|\nabla c(\bar{x})\|^2$, we have the contradiction

$$\rho\|\nabla c(x_r)\|^2 \geq \mu_r > \hat{\mu} > \rho\|\nabla c(\bar{x})\|^2 \;\text{ for all large } r.$$

For the special case of h convex and lower semicontinuous, we have from Theorem 6.5(c) that d_r with the properties (6.19) exists, for $\hat{\mu} = 0$. Uniqueness of d_r follows from strict convexity of h_{x_r,μ_r}. Validity of the chain rule, which is needed to obtain (6.13), follows as in the proof of Lemma 6.7. □

6.5. Manifold Identification. We next work toward the identification result. Consider a sequence of points $\{x_r\}$ in $\mathbb{R}^n$ converging to the critical point $\bar{x}$ of the composite function $h \circ c$, and let μ_r be a sequence of positive proximality parameters. Suppose now that the outer function h is partly smooth at the point $\bar{c} = c(\bar{x}) \in \mathbb{R}^m$ relative to some manifold $\mathcal{M} \subset \mathbb{R}^m$. Our aim is to find conditions guaranteeing that the update to the point $c(x_r)$ predicted by minimizing the proximal linearized objective h_{x_r,μ_r} lies on $\mathcal{M}$: in other words,

$$c(x_r) + \nabla c(x_r)d(x_r, \mu_r) \in \mathcal{M} \;\text{ for all large } r,$$

where $d(x_r, \mu_r)$ is the unique small critical point of h_{x_r,μ_r}. We would furthermore like to ensure that the "efficient projection" x_{new} resulting from this prediction, guaranteed by Theorem 6.6 (linear estimator improvement), satisfies $c(x_{\text{new}}) \in \mathcal{M}$.

To illustrate, we return to our ongoing example from Section 3, the case in which the outer function h is finite and polyhedral,

$$h(c) = \max_{i \in I}\{\langle h_i, c\rangle + \beta_i\},$$

for some given vectors $h_i \in \mathbb{R}^m$ and scalars β_i (see (3.1)). If $\bar{I}$ is the active index set corresponding to the point $\bar{c}$, then it is easy to check that h is partly smooth relative to the manifold

$$\mathcal{M} \;=\; \big\{c : \langle h_i, c\rangle + \beta_i = \langle h_j, c\rangle + \beta_j \text{ for all } i, j \in \bar{I}\big\}.$$

Our analysis requires one more assumption, in addition to those of Theorem 6.9. The basic criticality condition (1.3) requires the existence of a multiplier vector:

$$\partial h(\bar{c}) \cap \text{Null}(\nabla c(\bar{x})^*) \;\neq\; \emptyset.$$

We now strengthen this assumption slightly, to a "strict" criticality condition:

$$\text{ri}\big(\partial h(\bar{c})\big) \cap \text{Null}(\nabla c(\bar{x})^*) \;\neq\; \emptyset, \tag{6.20}$$

where ri denotes the relative interior of a convex set. This condition is related to the strict complementarity assumption in nonlinear programming. For h defined as above, since $\partial h(\bar{c}) = \text{conv}\{h_i : i \in \bar{I}\}$, we have

$$\text{ri}\big(\partial h(\bar{c})\big) \;=\; \Big\{\sum_{i \in \bar{I}} \lambda_i h_i : \sum_{i \in \bar{I}} \lambda_i = 1,\; \lambda > 0\Big\}.$$

Hence, the strict criticality condition (6.20) becomes the existence of a vector $\lambda \in \mathbb{R}^{\bar{I}}$ satisfying

$$\lambda > 0 \quad \text{and} \quad \sum_{i \in \bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{6.21}$$

The only change from the corresponding basic criticality condition (3.2) is that the condition $\lambda \geq 0$ has been strengthened to $\lambda > 0$, corresponding exactly to the extra requirement of strict complementarity in the nonlinear programming formulation (2.1).

Recall that the constraint qualification (6.11) implies the uniqueness of the multiplier vector $\bar{v}$, by Lemma 6.7. Assuming in addition the strict criticality condition (6.20), we then have

$$\bar{v} \in \operatorname{ri}\big(\partial h(\bar{c})\big) \cap \operatorname{Null}(\nabla c(\bar{x})^*).$$

We use the following result from Hare and Lewis [16], establishing a relationship between partial smoothness of functions and sets.

THEOREM 6.10. *([16, Theorem 5.1]) A function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ is partly smooth at a point $\bar{c} \in \mathbb{R}^m$ relative to a manifold $\mathcal{M} \subset \mathbb{R}^m$ if and only if the restriction $h|_{\mathcal{M}}$ is $\mathcal{C}^2$ around $\bar{c}$ and the epigraph* $\operatorname{epi} h$ *is partly smooth at the point $\big(\bar{c}, h(\bar{c})\big)$ relative to the manifold $\{\big(c, h(c)\big) : c \in \mathcal{M}\}$.*

We now prove a trivial modification of [16, Theorem 5.3].

THEOREM 6.11. *Suppose the function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ is partly smooth at the point $\bar{c} \in \mathbb{R}^m$ relative to the manifold $\mathcal{M} \subset \mathbb{R}^m$, and is prox-regular there. Consider a subgradient $\bar{v} \in \operatorname{ri} \partial h(\bar{c})$. Suppose the sequence $\{\hat{c}_r\} \subset \mathbb{R}^m$ satisfies $\hat{c}_r \to \bar{c}$ and $h(\hat{c}_r) \to h(\bar{c})$. Then $\hat{c}_r \in \mathcal{M}$ for all large r if and only if $\operatorname{dist}\big(\bar{v}, \partial h(\hat{c}_r)\big) \to 0$.*

Proof. The proof proceeds exactly as in [16, Theorem 5.3], except that instead of defining a function $g : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ by $g(c, r) = r$, we set $g(c, r) = r - c^T \bar{v}$. □

We can now prove our main identification result.

THEOREM 6.12. *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$, and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$ that is $\mathcal{C}^2$ around the point $\bar{x} \in \mathbb{R}^n$. Suppose that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and partly smooth there relative to the manifold $\mathcal{M}$. Suppose furthermore that the constraint qualification (6.11) and the strict criticality condition (6.20) both hold for the composite function $h \circ c$ at $\bar{x}$. Then there exist constants $\hat{\mu}, \gamma \geq 0$ with the following property. Given any sequence $\{\mu_r\}$ with $\mu_r > \hat{\mu}$ for all r, and any sequence $x_r \to \bar{x}$ such that $\mu_r |x_r - \bar{x}| \to 0$, the local minimizer d_r of h_{x_r, μ_r} defined in Theorem 6.9 satisfies, for all large r, the condition*

$$c(x_r) + \nabla c(x_r) d_r \in \mathcal{M}, \tag{6.22}$$

and also the inequalities

$$|x_r^{\text{new}} - (x_r + d_r)| \leq \gamma |d_r|^2 \quad \text{and} \quad h\big(c(x_r^{\text{new}})\big) \leq h\big(c(x_r) + \nabla c(x_r) d_r\big) + \gamma |d_r|^2, \tag{6.23}$$

hold for some point x_r^{new} with $c(x_r^{\text{new}}) \in \mathcal{M}$.

In the special case when $h : \mathbb{R}^m \to (-\infty, +\infty]$ is convex and lower semicontinuous function, the result holds with $\hat{\mu} = 0$.

Proof. Theorem 6.9 implies $d_r \to 0$, so

$$\hat{c}_r = c(x_r) + \nabla c(x_r) d_r \to \bar{c}.$$

The theorem also shows $h(\hat{c}_r) \to h(\bar{c})$, and furthermore that there exist multiplier vectors $v_r \in \partial h(\hat{c}_r)$ satisfying

$$v_r \to \bar{v} \in \operatorname{ri} \partial h(\bar{c}).$$

Since

$$\operatorname{dist}\big(\bar{v}, \partial h(\hat{c}_r)\big) \leq |\bar{v} - v_r| \to 0,$$

we can apply Theorem 6.11 to obtain property (6.22).

Let us now define a function $h_{\mathcal{M}} \colon \mathbb{R}^m \to \bar{\mathbb{R}}$, agreeing with h on the manifold $\mathcal{M}$ and taking the value $+\infty$ elsewhere. By partial smoothness, $h_{\mathcal{M}}$ is the sum of a smooth function and the indicator function of $\mathcal{M}$, and hence $\partial^\infty h_{\mathcal{M}}(\bar{c}) = N_{\mathcal{M}}(\bar{c})$. Partial smoothness also implies $\operatorname{par}\big(\partial h(\bar{c})\big) = N_{\mathcal{M}}(\bar{c})$. We can therefore rewrite the constraint qualification (6.11) in the form

$$\partial^\infty h_{\mathcal{M}}(\bar{c}) \cap \operatorname{Null}\big(\nabla c(\bar{x})^*\big) = \{0\}.$$

This condition allows us to apply Theorem 6.6 (linear estimator improvement), with the function $h_{\mathcal{M}}$ replacing the function h, to deduce the existence of the point x_r^{new}, as required. □

7. A Proximal Algorithm and its Properties. We now describe a simple first-order algorithm that manipulates the proximality parameter μ in (4.1) to achieve a "sufficient decrease" in h at each iteration. We follow up with some results concerning the global convergence behavior of this method and its ability to identify the manifold $\mathcal{M}$ of Section 6.5.

Algorithm ProxDescent
Define constants $\tau > 1$, $\sigma \in (0,1)$, and $\mu_{\min} > 0$;
Choose x_0, $\mu_0 \geq \mu_{\min}$;
Set $\mu \leftarrow \mu_0$;
for $k = 0, 1, 2, \ldots$
    Set accept $\leftarrow$ false;
    while not accept
        Find a local minimizer d_k of (4.1) with $x = x_k$
            such that $h_{x_k,\mu}(d_k) < h_{x_k,\mu}(0)$;
        if no such d exists
            terminate with $\bar{x} = x_k$;
        end (if)
        Derive x_k^+ from $x_k + d_k$ (by an efficient projection and/or
                other enhancements);
        if $h\big(c(x_k)\big) - h\big(c(x_k^+)\big) \geq \sigma \left[h\big(c(x_k)\big) - h(c(x_k) + \nabla c(x_k) d_k)\right]$
                and $|x_k^+ - (x_k + d_k)| \leq \frac{1}{2}|d_k|$
            $x_{k+1} \leftarrow x_k^+$;
            $\mu_k \leftarrow \mu$;
            $\mu \leftarrow \max(\mu_{\min}, \mu/\tau)$;
            accept $\leftarrow$ true;
        else
            $\mu \leftarrow \tau\mu$;
        end (if)
    end (while)
end (for).

We are not overly specific about the the derivation of x_k^+ from $x_k + d_k$, but we assume that the "efficient projection" technique that is the basis of Theorem 6.6 is used when possible. Lemma 6.8 indicates that for μ sufficiently large and x near a critical point $\bar{x}$ of $h \circ c$, it is indeed possible to find a local solution d of (4.1) which satisfies $h_{x,\mu}(d) < h_{x,\mu}(0)$ as required by the algorithm, and which also satisfies the conditions of Theorem 6.6. Lemma 7.2 below shows further that the new point x_k^+ satisfies the acceptance tests in the algorithm. However, Lemma 7.2 is more general in that it also gives conditions for acceptance of the step when x_k is *not* in a neighborhood of a critical point of $h \circ c$.

The framework also allows x_k^+ to be improved further. For example, we could use higher-order derivatives of c to take a further step along the manifold of h identified by the subproblem (4.1), analogous to an "EQP step" in nonlinear programming, and reset x_k^+ accordingly if this step produces a reduction in $h \circ c$. We discuss this point further at the end of the section.

The main result in this section — Theorem 7.4 — specifies conditions under which Algorithm ProxDescent does not have nonstationary accumulation points. We start with a technical result that in the neighborhood of a non-critical point $\bar{x}$ and for bounded μ, the steps d do not become too short.

LEMMA 7.1. *Consider a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Let $\bar{x}$ be such that: c is $\mathcal{C}^1$ near $\bar{x}$; h is finite at the point $\bar{c} = c(\bar{x})$ and subdifferentially regular there; the transversality condition (6.3) holds; but the criticality condition (1.3) is* not *satisfied. Then given any constant $\mu_{\max} \geq 0$, there exists a quantity $\epsilon > 0$ such that for any sequence $x_r \to \bar{x}$ with $h\big(c(x_r)\big) \to h\big(c(\bar{x})\big)$, and any sequence $\mu_r \in [0, \mu_{\max}]$, any sequence of critical points d_r of h_{x_r,μ_r} satisfying $h_{x_r,\mu_r}(d_r) \leq h_{x_r,\mu_r}(0)$ must also satisfy $\liminf_r |d_r| \geq \epsilon$.*

Proof. If the result failed, there would exist sequences x_r, μ_r, and d_r as above except that $d_r \to 0$. Noting that $h(c(x_r)+\nabla c(x_r)d_r) \to h\big(c(\bar{x})\big)$ (using lower semicontinuity and the fact that the left-hand side is dominated by $h\big(c(x_r)\big)$, which converges to $h(\bar{c})$), we have that

$$\partial^\infty h\big(c(x_r) + \nabla c(x_r)d_r\big) \cap \operatorname{Null}(\nabla c(x_r)^*) \;=\; \{0\},$$

for all r sufficiently large. (If this were not true, we could use an outer semicontinuity argument based on [34, Theorem 8.7] to deduce that $\partial^\infty h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*)$ is nonempty, thus violating the transversality condition (6.3).) Hence, we can apply the chain rule and deduce that there are multiplier vectors v_r such that (6.13) is satisfied, that is,

$$\begin{aligned} 0 &= \nabla c(x_r)^* v_r + \mu_r d_r, \\ v_r &\in \partial h\big(c(x_r) + \nabla c(x_r)d_r\big), \end{aligned}$$

for all sufficiently large r. If the sequence $\{v_r\}$ is unbounded, we can assume without loss of generality that $|v_r| \to \infty$. Any limit point of the sequence $v_r/|v_r|$ would be a unit vector in the set $\partial^\infty(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*)$, contradicting (6.3). Hence, the sequence $\{v_r\}$ is bounded, so by taking limits in the conditions above and using $\mu_r d_r \to 0$ and outer semicontinuity of $\partial h(c)$ at $\bar{c}$, we can identify a vector $\bar{v}$ such that $\bar{v} \in \partial h(\bar{c}) \cap \operatorname{Null}(\nabla c(\bar{x})^*)$. Using the chain rule and subdifferential regularity, this contradicts non-criticality of $\bar{x}$. □

The next result makes use of the efficient projection mechanism of Theorem 6.6. When the conditions of this theorem are satisfied, we show that the Algorithm Prox-

Descent can perform the projection to obtain the point x_k^+ in such a way that (6.10) is satisfied. We thus have the following result.

LEMMA 7.2. *Consider the constant $\sigma \in (0,1)$, a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$, and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$ that is $\mathcal{C}^2$ around a point $\bar{x} \in \mathbb{R}^n$. Assume that h lower semicontinuous and finite at $\bar{c} = c(\bar{x})$ and that transversality condition (6.3) holds at $\bar{x}$. Then there exist constants $\tilde{\mu} > 0$ and $\tilde{\delta} > 0$ with the following property: For any $x \in B_{\tilde{\delta}}(\bar{x})$, $d \in B_{\tilde{\delta}}(0)$, and $\mu \geq \tilde{\mu}$ such that*

$$h_{x,\mu}(d) \leq h_{x,\mu}(0), \qquad |h(c(x) + \nabla c(x)d) - h\big(c(\bar{x})\big)| \leq \tilde{\delta}, \tag{7.1}$$

there is a point $x^+ \in \mathbb{R}^n$ such that

$$h\big(c(x)\big) - h\big(c(x^+)\big) \geq \sigma\big[h\big(c(x)\big) - h\big(c(x) + \nabla c(x)d\big)\big], \tag{7.2a}$$

$$|x^+ - (x+d)| \leq \tfrac{1}{2}|d|. \tag{7.2b}$$

Proof. Define δ and γ as in Theorem 6.6 and set $\tilde{\delta} = \min\big(\delta, 1/(2\gamma)\big)$. By applying Theorem 6.6, we obtain a point x^+ (denoted by $x_{\rm new}$ in the earlier result) for which $|x^+ - (x+d)| \leq \gamma|d|^2 \leq \frac{1}{2}|d|$ (thus satisfying (7.2b)) and $h\big(c(x^+)\big) \leq h(c(x) + \nabla c(x)d) + \gamma|d|^2$. Also note that because of $h_{x,\mu}(d) \leq h_{x,\mu}(0)$, we have

$$h(c(x) + \nabla c(x)d) + \frac{\mu}{2}|d|^2 \leq h\big(c(x)\big)$$

and hence

$$|d|^2 \leq \frac{2}{\mu}\left[h\big(c(x)\big) - h(c(x) + \nabla c(x)d)\right].$$

We therefore have

$$\begin{aligned} h\big(c(x)\big) - h\big(c(x^+)\big) &\geq h\big(c(x)\big) - h(c(x) + \nabla c(x)d) - \gamma|d|^2 \\ &\geq \left[h\big(c(x)\big) - h(c(x) + \nabla c(x)d)\right]\left(1 - \frac{2\gamma}{\mu}\right). \end{aligned}$$

By choosing $\tilde{\mu}$ large enough that $1 - 2\gamma/\tilde{\mu} > \sigma$, we obtain (7.2a). □

We also need the following elementary lemma.

LEMMA 7.3. *For any constants $\tau > 1$ and $\rho > 0$ and any positive integer t, we have*

$$\min\Big\{\sum_{i=1}^t \alpha_i^2 \tau^i : \sum_{i=1}^t \alpha_i \geq \rho,\ \alpha \in \mathbb{R}_+^t\Big\} \quad > \quad \rho^2(\tau - 1).$$

Proof. By scaling, we can suppose $\rho = 1$. Clearly the optimal solution of this problem must lie on the hyperplane $H = \{\alpha : \sum_i \alpha_i = 1\}$. The objective function is convex, and its gradient at the point $\bar{\alpha} \in H$ defined by

$$\bar{\alpha}_i = \frac{\tau^{1-i} - \tau^{-i}}{1 - \tau^{-t}} > 0$$

is easily checked to be orthogonal to H. Hence $\bar{\alpha}$ is optimal, and the corresponding optimal value is easily checked to be strictly large than $\tau - 1$. □

In the following result, the assumptions on h, c, and $\bar{x}$ allow us to apply both Lemmas 7.1 and 7.2.

THEOREM 7.4. *Consider a constant $\sigma \in (0,1)$, a function $h\colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and a map $c\colon \mathbb{R}^n \to \mathbb{R}^m$. Let the point $\bar{x} \in \mathbb{R}^n$ be such that c is $\mathcal{C}^2$ near $\bar{x}$; h is subdifferentially regular at the point $\bar{c} = c(\bar{x})$; the transversality condition (6.3) holds; and the criticality condition (1.3) is* not *satisfied. Then the pair $\big(\bar{x}, h(\bar{c})\big)$ cannot be an accumulation point of the sequence $\big(x_k, h\big(c(x_k)\big)\big)$ generated by Algorithm ProxDescent.*

Proof. Suppose for contradiction that $\big(\bar{x}, h(\bar{c})\big)$ is an accumulation point. Since the sequence $\{h\big(c(x_r)\big)\}$ generated by the algorithm is monotonically decreasing, we have $h\big(c(x_r)\big) \downarrow h(\bar{c})$. By the acceptance test in the algorithm and the definition of $h_{x,\mu}$ in (4.1), we have that

$$\begin{aligned} h\big(c(x_{r+1})\big) &\le h\big(c(x_r)\big) - \sigma\big[h\big(c(x_r)\big) - h\big(c(x_r) + \nabla c(x_r)d_r\big)\big] \\ &\le h\big(c(x_r)\big) - \sigma\frac{\mu_r}{2}|d_r|^2. \end{aligned} \tag{7.3}$$

Using this inequality, we have

$$\begin{aligned} h\big(c(x_0)\big) - h\big(c(\bar{x})\big) &\ge \sum_{r=0}^{\infty} h\big(c(x_r)\big) - h\big(c(x_{r+1})\big) \\ &\ge \frac{\sigma}{2}\sum_{r=1}^{\infty} \mu_r |d_r|^2 \\ &\ge \frac{\sigma}{2}\mu_{\min}\sum_{r=1}^{\infty} |d_r|^2, \end{aligned}$$

which implies that $d_r \to 0$. Further, we have that

$$\begin{aligned} &|h(c(x_r)+\nabla c(x_r)d_r) - h(\bar{c})| \\ &\qquad \le \big[h\big(c(x_r)\big) - h(c(x_r) + \nabla c(x_r)d_r)\big] + \big[h\big(c(x_r)\big) - h(\bar{c})\big] \\ &\qquad \le \sigma^{-1}\big[h\big(c(x_r)\big) - h\big(c(x_{r+1})\big)\big] + \big[h\big(c(x_r)\big) - h(\bar{c})\big] \to 0. \end{aligned} \tag{7.4}$$

Because $\bar{x}$ is an accumulation point, we can define a subsequence of indices r_j, $j = 0, 1, 2, \ldots$ such that $\lim_{j\to\infty} x_{r_j} = \bar{x}$. The corresponding sequence of regularization parameters μ_{r_j} must be unbounded, since otherwise we could set $\mu_{\max}$ in Lemma 7.1 to be an upper bound on μ_{r_j}, and deduce that the sequence $|d_{r_j}|$ is bounded away from zero, which contradicts $d_r \to 0$. Defining $\tilde{\mu}$ and $\tilde{\delta}$ as in Lemma 7.2, we can assume without loss of generality that $\mu_{r_j} > \tau\tilde{\mu}$ and $\mu_{r_{j+1}} > \mu_{r_j}$ for all j. Moreover, since $x_{r_j} \to \bar{x}$ and $d_{r_j} \to 0$, and using (7.4), we can assume that

$$x_{r_j} \in B_{\tilde{\delta}/2}(\bar{x}), \qquad \text{for } j = 0, 1, 2, \ldots, \tag{7.5a}$$

$$d_r \in B_{\tilde{\delta}}(0), \qquad \text{for all } r > r_0, \tag{7.5b}$$

$$|h(c(x_r) + \nabla c(x_r)d_r) - h(\bar{c})| \le \tilde{\delta}, \quad \text{for all } r > r_0. \tag{7.5c}$$

The value of μ cannot be increased in the inner iteration of Algorithm ProxDescent at iteration r_j. We verify this claim by noting that because of (7.5), Lemma 7.2 tells us that the previously tried value of μ, namely $\mu_{r_j}/\tau > \tilde{\mu}$, would have been accepted by the algorithm had it tried to increase μ during iteration r_j. We define k_j to be the latest iteration prior to r_{j+1} at which μ was increased, in the inner iteration of

Algorithm ProxDescent. Note that such an iteration k_j exists, because $\mu_{r_{j+1}} > \mu_{r_j}$, so the value of μ must have been increased during *some* intervening iteration. Moreover, we have $r_j < k_j < r_{j+1}$. Since no increases of μ were performed internally during iterations $k_j + 1, \ldots, r_{j+1}$, the value of μ used at these each iterations was the first one tried, which was a factor τ^{-1} of the value from the previous iteration. That is,

$$\tau\tilde{\mu} < \mu_{r_{j+1}} = \tau^{-1}\mu_{r_{j+1}-1} = \tau^{-2}\mu_{r_{j+1}-2} = \ldots = \tau^{k_j - r_{j+1}}\mu_{k_j}. \tag{7.6}$$

Since the previous value of μ tried at iteration k_j, namely μ_{k_j}/τ, was rejected, we can conclude from Lemma 7.2 that $|x_{k_j} - \bar{x}| > \tilde{\delta}$. To see this, note that all the other conditions of Lemma 7.2 are satisfied by this value of μ, that is, $\mu_{k_j}/\tau \geq \tilde{\mu}$, $d_{k_j} \in B_{\tilde{\delta}}(0)$ (because of (7.5b)), and $\big|h(c(x_{r_j}) + \nabla c(x_{r_j})d_{r_j}) - h(\bar{c})\big| \leq \tilde{\delta}$ (because of (7.5c)). Recalling that $|x_{r_{j+1}} - \bar{x}| < \tilde{\delta}/2$, and noting from the acceptance criteria in Algorithm ProxDescent that $|x_{k+1} - x_k| \leq |x_{k+1} - (x_k + d_k)| + |d_k| \leq (3/2)|d_k|$, we have that

$$\frac{1}{2}\tilde{\delta} < |x_{r_{j+1}} - x_{k_j}| \leq \sum_{k=k_j}^{r_{j+1}-1} |x_{k+1} - x_k| = \frac{3}{2}\sum_{k=k_j}^{r_{j+1}-1} |d_k|. \tag{7.7}$$

To bound the decrease in objective function over the steps from x_{k_j} to $x_{r_{j+1}}$, we have from the acceptance condition and (7.6) that

$$\begin{aligned} h\big(c(x_{k_j})\big) - h\big(c(x_{r_{j+1}})\big) &= \sum_{k=k_j}^{r_{j+1}-1} h\big(c(x_k)\big) - h\big(c(x_{k+1})\big) \\ &\geq \frac{\sigma}{2}\sum_{k=k_j}^{r_{j+1}-1} \mu_k |d_k|^2 \\ &= \frac{\sigma}{2}\mu_{r_{j+1}}\tau^{-1}\sum_{k=k_j}^{r_{j+1}-1} \tau^{r_{j+1}-k}|d_k|^2. \end{aligned}$$

To obtain a lower bound on the final summation, we apply Lemma 7.3 with $\rho = \tilde{\delta}/3$ (from (7.7)) and $t = r_{j+1} - k_j \geq 1$ to obtain

$$h\big(c(x_{k_j})\big) - h\big(c(x_{r_{j+1}})\big) \geq \frac{\sigma}{2}\mu_{r_{j+1}}\tau^{-1}\Big(\frac{\tilde{\delta}}{3}\Big)^2(\tau - 1) \geq \frac{\sigma\tilde{\mu}\tilde{\delta}^2(\tau-1)}{18} > 0,$$

where we have used $\mu_{r_{j+1}} > \tau\tilde{\mu}$. Since this finite decrease happens for every index $j = 1, 2, \ldots$, we obtain a contradiction from the usual telescoping sum argument. □

To illustrate the idea of identification, we state a simple manifold identification result for the case when the function h is convex and finite.

THEOREM 7.5. *Consider a function $h : \mathbb{R}^m \to \mathbb{R}$, a map $c : \mathbb{R}^n \to \mathbb{R}^m$, and a point $\bar{x} \in \mathbb{R}^n$ that is critical for $h \circ c$. Suppose that c is C^2 near $\bar{x}$, and that h is convex and continuous on* $\operatorname{dom} h$ *near $\bar{c} := c(\bar{x})$. Suppose in addition that h is partly smooth at $\bar{c}$ relative to the manifold $\mathcal{M}$. Finally, assume that the constraint qualification (6.11) and the strict criticality condition (6.20) both hold for the composite function $h \circ c$ at $\bar{x}$.*

Then if Algorithm ProxDescent generates a sequence $x_r \to \bar{x}$, we have that $c(x_r) + \nabla c(x_r)d_r \in \mathcal{M}$ for all r sufficiently large.

Proof. Note that h, c, and $\bar{x}$ satisfy the assumptions of Theorem 6.12, with $\hat{\mu} = 0$. To apply Theorem 6.12 and thus prove the result, we need to show only that $\mu_r |x_r - \bar{x}| \to 0$. In fact, we show that $\{\mu_r\}$ is bounded, so that this estimate is satisfied trivially.

Using Lemma 7.2, we have that the step acceptance condition of Algorithm Prox-Descent is satisfied at x_r for all $\mu \geq \tilde{\mu}$. It follows that for all r sufficiently large, we have in fact that $\mu_r \leq \tau\tilde{\mu}$, which leads to the desired result. □

To enhance the step d obtained from (4.1), we might try to incorporate second-order information inherent in the structure of the subdifferential ∂h at the new value of c predicted by the linearized subproblem. Knowledge of the subdifferential $\partial h\big(c_{\rm pred}(x)\big)$ allows us in principle to compute the tangent space to $\mathcal{M}$ at $c_{\rm pred}(x)$. We could then try to "track" $\mathcal{M}$ using second-order information, since both the map c and the restriction of the function h to $\mathcal{M}$ are $\mathcal{C}^2$.

REFERENCES

[1] J. V. Burke, *On the identification of active constraints II: The nonconvex case*, SIAM Journal on Numerical Analysis, 27 (1990), pp. 1081–1102.

[2] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM Journal on Numerical Analysis, 25 (1988), pp. 1197–1211.

[3] R. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *On the convergence of successive linear-quadratic programming algorithms*, SIAM Journal on Optimization, 16 (2005), pp. 471–489.

[4] J.-F. Cai, E. Candès, and Z. Shen, *A singular valur thresholding algorithm for matrix completion*, technical report, Applied and Computational Mathematics, California Institute of Technology, September 2008.

[5] E. Candès and B. Recht, *Exact matrix completion via convex optimization*, Technical Report, California Institute of Technology, 2008.

[6] E. J. Candès, *Compressive sampling*, in Proceedings of the International Congress of Mathematicians, Madrid, 2006.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

[8] P. Combettes and T. Pennanen, *Proximal methods for cohypomonotone operators*, SIAM Journal on Control and Optimization, 43 (2004), pp. 731–742.

[9] A. Daniilidis, W. Hare, and J. Malick, *Geometrical interpretation of the proximal-type algorithms in structured optimization problems*, Optimization, 55 (2006).

[10] A. V. Dmitruk and A. Y. Kruger, *Metric regularity and systems of generalized equations*, Journal of Mathematical Analysis and Applications, 342 (2008), pp. 864–873.

[11] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radiius of metric regularity*, Transactions of the American Mathematical Society, 355 (2003), pp. 493–517.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics, 32 (2004), pp. 407–499.

[13] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Mathematical Programming, 43 (1989), pp. 235–256.

[14] M. P. Friedlander, N. I. M. Gould, S. Leyffer, and T. S. Munson, *A filter active-set trust-region method*, Preprint ANL/MCS-P1456-0907, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne IL 60439, September 2007.

[15] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain nonconvex minimization problems*, International Journal of Systems Science, 12 (1981), pp. 989–1000.

[16] W. Hare and A. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.

[17] A. Iusem, T. Pennanen, and B. Svaiter, *Inexact variants of the proximal point algorithm without monotonicity*, SIAM Journal on Optimization, 13 (2003), pp. 1080–1097.

[18] S. Jokar and M. E. Pfetsch, *Exact and approximate sparse solutions of underdetermined linear equations*, ZIB-Report 07-05, ZIB, March 2007.

[19] A. Kaplan and R. Tichatschke, *Proximal point methods and nonconvex optimization*, Journal of Global Optimization, 13 (1998), pp. 389–406.

[20] C. Lemaréchal, F. Oustry, and C. Sagastizabal, *The u-lagrangian of a convex function*, Transactions of the American Mathematical Society, 352 (2000), pp. 711–729.
[21] A. Levy, *Lipschitzian multifunctions and a Lipschitzian inverse mapping theorem*, Mathematics of Operations Research, 26 (2001), pp. 105–118.
[22] A. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13 (2003), pp. 702–725.
[23] O. L. Mangasarian, *Minimum-support solutions of polyhedral concave programs*, Optimization, 45 (1999), pp. 149–162.
[24] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.
[25] R. Mifflin and C. Sagastizábal, *A VU-algorithm for convex minimization*, Mathematical Programming, Series B, 104 (2005), pp. 583–608.
[26] S. A. Miller and J. Malick, *Newton methods for nonsmooth convex minimization: Connections among U-Lagrangian, Reimannian Newton, and SQP methods*, Mathematical Programming, Series B, 104 (2005), pp. 609–633.
[27] B. Mordukhovich, *Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications*, Springer, New York, 2006.
[28] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Mathematics of Operations Research, 27 (2002), pp. 170–191.
[29] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Transactions of the American Mathematical Society, 348 (1996), pp. 1805–1838.
[30] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Transactions of the American Mathematical Society, 352 (2000), pp. 5231–5249.
[31] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum-rank solutions of matrix equations via nuclear norm minimiation*, tech. report, Massachusetts Institute of Technology, 2007.
[32] R. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.
[33] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
[34] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer, 1998.
[35] C. Sagastizábal and R. Mifflin, *Proximal points are on the fast track*, Journal of Convex Analysis, 9 (2002), pp. 563–579.
[36] W. Shi, G. Wahba, S. J. Wright, K. Lee, R. Klein, and B. Klein, *LASSO-Patternsearch algorithm with application to opthalmology data*, Statistics and its Interface, 1 (2008), pp. 137–153.
[37] J. Spingarn, *Submonotone mappings and the proximal point algorithm*, Numerical Functional Analysis and Optimization, 4 (1981/82), pp. 123–150.
[38] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society B, 58 (1996), pp. 267–288.
[39] S. J. Wright, *Convergence of an inexact algorithm for composite nonsmooth optimization*, IMA Journal of Numerical Analysis, 9 (1990), pp. 299–321.
[40] ———, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079.
[41] Y. Yuan, *Conditions for convergence of a trust-region method for nonsmooth optimization*, Mathematical Programming, 31 (1985), pp. 220–228.
[42] ———, *On the superlinear convergence of a trust region algorithm for nonsmooth optimization*, Mathematical Programming, 31 (1985), pp. 269–285.
[43] H. H. Zhang, J. Ahn, X. Lin, and C. Park, *Gene selection using support vector machines with non-convex penalty*, Bioinformatics, 22 (2006), pp. 88–95.